

Intelligence artificielle : Pierre pensante ou pierre philosophale ?

Esquisse d'une Théorie Méréologique
du Jugement Intuitif

Membres du jury :

Peter Verdée, co-promoteur
Tom Dedeurwaerdere, co-promoteur
Bruno Leclercq, lecteur
Bart Van Kerkhove, lecteur
David Watson, lecteur
Mark Hunyadi, président

Thèse réalisée par
Gwenaël G. C. Laurent

en vue de l'obtention du grade de
Doctorat en philosophie

Bien qu'elles fissent plusieurs choses aussi bien ou peut-être mieux qu'aucun de nous, [ces machines] manqueraient infailliblement en quelques autres, par lesquelles on découvrirait qu'elles n'agiraient pas par connaissance, mais seulement par la disposition de leurs organes.

DESCARTES

Table des matières

Introduction	11
Motivations éthiques	15
Un problème philosophique	17
Programmation top-down	19
Programmation bottom-up	22
Les Réseaux de neurones	23
Les Transformeurs	24
Le Jugement intuitif	25
Le QI artificiel	29
Plan de la thèse	32
Partie 1 : Le Problème du jugement en philosophie	33
Partie 2 : Révision du concept de jugement	36
Partie 3 : Le Jugement intuitif face à la Machine	42
Partie 4 : Vers une théorie générale de l'esprit humain	45
Le Problème du jugement en philosophie	
De l'analytique au synthétique	53
1 Le Jugement chez Platon :	
L'Origine du problème	53
1.1 Socrate vs Protagoras :	
Les Monologues de l'âme	57
1.1.1 Le Jugement d'expérience	59
1.1.2 Le Jugement vrai sans justification	59
1.1.3 Le Jugement vrai muni de raison	61
1.2 La Contemplation du Beau	64
1.2.1 L'Analogie de la ligne	65
1.2.2 Les Pathèmata	66
2 Le Jugement chez Kant :	
La Genèse de l'Entendement	71
2.1 L'Abstraction intuitive	72

2.1.1	Le Jugement synthétique <i>a priori</i>	75
2.1.2	Les Limites du platonisme	79
2.2	Kant vs Platon : Le Procès des Idées	82
2.2.1	La Métaphore du colombier	83
2.2.2	La Connaissance <i>a priori</i>	87
2.2.3	La Révolution copernicienne	91
2.3	La Sensibilité : l'Homme qui sent	92
2.4	L'Entendement : L'Homme qui pense	95
2.4.1	Kant vs François Chollet	96
2.4.2	La Dédution objective	100
2.4.3	La Dédution subjective	101
2.4.4	Le Schématisme kantien	102
2.5	La Raison : L'Homme qui pense qu'il pense	104
2.6	La Beauté sans vérité	107
2.6.1	L'Importance de la première critique	112
3	Le Jugement artificiel :	
	La Machine peut-elle juger ?	115
3.1	La Machine manque-t-elle d'intuition spatiale ?	121
3.2	La Machine manque-t-elle d'intuition temporelle ?	125
3.2.1	Le Modèle des saccades	126
3.2.2	Objection : la Chronostase auditive	127
3.2.3	Hypothèse : perception différentielle	130
3.2.4	Temps subjectif vs temps objectif	134
3.3	La Machine manque-t-elle de raison contemplative ?	136
3.4	La Machine manque-t-elle d'Entendement ?	139
3.4.1	Lacunes au niveau des catégories	142
3.4.2	Logique générale vs transcendantale	144
3.5	La Machine est-elle un Je ?	147
3.6	La Machine pense-t-elle différemment ?	148
	Révision du concept de jugement	
	Vers une théorie Méréologique	155
4	L'Oubli de l'Entendement :	
	Mort du jugement kantien	155
4.1	La Crise des fondements des mathématiques	161
4.2	Le Livre de Dieu et les Mondes parallèles	166
4.2.1	Géométries sphériques et elliptiques	170
4.2.2	Géométries hyperboliques	172
4.3	Une Géométrie imaginaire	174
4.3.1	Le Constructivisme de Lobatchevski	176
4.3.2	Le Consensualisme de Poincaré	177

4.4	Kant se retourne dans la tombe	178
4.4.1	L'Espace est-il une chose en soi ?	178
4.4.2	L'Espace est-il une construction ?	180
4.4.3	Les Mathématiques sont-elles synthétiques <i>a priori</i> ?	181
5	La Rationalité numérique :	
	Naissance de l'ordinateur	185
5.1	Le Rêve d'un Calculateur universel	187
5.2	Les Théories de la démontrabilité	190
5.2.1	Le Paradoxe du barbier	191
5.2.2	Le Paradoxe de Cantor	193
5.2.3	Contourner les paradoxes	195
5.2.4	Les Théorèmes de Gödel	198
5.2.5	La Vérité est une passoire	201
5.3	Les Théories de la calculabilité	205
5.3.1	Formalisme de la Machine de Turing	208
5.3.2	Le Problème de l'arrêt	217
5.3.3	Le Théorème de Rice	220
5.3.4	Oracles et cybernétique	222
6	Approche Top Down :	
	Les Systèmes Experts	229
6.1	The General Problem Solver	236
6.2	Logique propositionnelle	240
6.2.1	Alphabet	241
6.2.2	Grammaire	242
6.2.3	Sémantique	243
6.2.4	Formules	245
6.2.5	Axiomes	245
6.3	Les Systèmes Experts	248
6.4	Forme normale conjonctive	251
6.5	Problème de satisfaisabilité (SAT)	254
6.5.1	Complexité de 2-SAT	256
6.5.2	Complexité de 3-SAT	258
6.5.3	Complexité de CNF-SAT	259
6.5.4	Complexité spatiale	260
6.6	Problèmes NP-complets	260
6.7	Critique épistémologique :	
	Entre platonisme et logicisme	266
7	Approche Bottom-Up :	
	Les Réseaux de neurones	271
7.1	Le Neurone biologique vs formel	272
7.1.1	Le Neurone formel	275

7.2	La Règle de Hebb : qui se ressemble s'assemble !	278
7.2.1	Neurone isolé	279
7.2.2	Réseau à une couche	284
7.2.3	Réseaux à couches multiples	284
7.3	Deep Learning : Mythe ou réalité ?	286
7.3.1	Programmation génétique	290
7.3.2	Apprentissage non supervisé	292
7.4	Critique épistémologique :	
	Entre instinct et raisonnement	295
7.4.1	Le Talon d'Achille des réseaux de neurones	296
7.4.2	Le Réseau de neurones vs le cerveau biologique	300
8	L'Angle mort de l'IA :	
	Esquisse d'une théorie holistique du jugement	307
8.1	Les Concepts intuitifs	312
8.2	Une théorie tricamérale de la cognition	316
8.2.1	La Rationalité : les concepts abstraits	317
8.2.2	La Sensation : les percepts statistiques	319
8.2.3	La Perception : les concepts intuitifs	321
8.3	Critique de la séquentialité :	
	L'intuition du Tout	326
8.4	Critique de l'univocité :	
	L'intuition de l'Infini	328
	Le Jugement intuitif face à la Machine	
	Extension de la machine de Turing	337
9	Le Dilemme de l'intuition	337
9.1	L'Intuition de l'Infini et l'intuition du Tout	340
9.2	L'Objection d'Alan Turing	345
9.3	L'Intuition selon Hubert Dreyfus	348
9.3.1	Objections à la position de Dreyfus	353
9.3.2	Sortir du statu quo	355
9.4	Première solution : le Mystérianisme	356
9.4.1	Ignoramus et ignorabimus	357
9.4.2	La posture de McGinn	358
9.4.3	Principales objections	361
9.4.4	L'Intuition est-elle un mystère ?	362
9.5	Deuxième solution : l'Intuition intellectuelle	364
9.5.1	La Contemplation des Idées	365
9.5.2	La Critique kantienne	369

10 La Machine de Chuck Norris	373
10.1 Fonctions et Modèles	375
10.1.1 Effet SNARC : structure intuitive de l'arithmétique	376
10.1.2 Vers un modèle de calcul transfini	379
10.1.3 L'intuition de l'Infini face aux problèmes indécidables	381
10.1.4 L'intuition du Tout face aux problèmes NP-complets	382
10.1.5 Réponses à une objection préliminaire	383
10.2 La Machine de Turing transfinie	385
10.2.1 Définition formelle	388
10.2.2 Restrictions et extensions d'une machine transfinie	394
10.2.3 Puissance de calcul	396
10.2.4 Représentation minimale	402
10.2.5 Machine à k bandes	405
10.3 État de l'Art	407
10.3.1 Machine à oracles hiérarchisés	407
10.3.2 Machine de Turing continue	410
10.3.3 Machine de Turing Temporelle Infinie	417
10.4 Machine universelle et problème de l'arrêt généralisé	420
11 Problèmes NP-complets généralisés	429
11.1 Les Problèmes NP-complets classiques	432
11.2 Le Problème 3-SAT _{N₀}	435
11.3 Représentation géométrique du problème	436
11.4 Méthode de résolution	442
11.5 Programme pour une machine transfinie	449
11.6 Stratégies concernant $\mathcal{P} = \mathcal{NP}$	451
Vers une théorie générale de l'esprit humain	
Ce que disent les neurosciences	459
12 Le Problème Corps-Esprit :	
Mystérianisme perspectiviste	459
12.1 L'Insoutenable unité du corps et de l'esprit	462
12.2 Hypothèse 1 : Conscience vive et conscience morte	465
12.3 Hypothèse 2 : Le Paradoxe de la conscience	468
12.3.1 La Conscience est irréprésentable	470
12.3.2 Chiasmes externalistes et internalistes	472
12.3.3 Liens avec les paradoxes logiques	475
12.4 Hypothèse 3 : Monisme à double aspect	482
12.4.1 Première analogie : L'Espace et les corps étendus	486
12.4.2 Deuxième analogie : Les Théorèmes d'incomplétude	493
12.4.3 Troisième analogie : Le Principe de complémentarité	494
12.4.4 Quatrième analogie : L'Hypersphère de Dante	497

13 Critique de l'encéphalocentrisme	503
13.1 La Conscience est-elle dans le cerveau?	505
13.2 Ségrégation fonctionnelle et plasticité neuronale	508
13.3 Spinozisme neuroscientifique	511
14 La Théorie darwinienne de Gerald Edelman	517
14.1 Critique edelmanienne du dualisme	518
14.2 Critique edelmanienne du computationnalisme	521
14.3 Théorie de la sélection des groupes neuronaux (TSGN)	524
14.4 Théorie edelmanienne de la conscience	529
14.5 Chiasmes métaphysiques et principes épistémiques	530
15 La Théorie Intégrée de l'Information	537
15.1 Stratégie de l'équivalence structurelle	540
15.2 Axiomes de l'expérience phénoménologique	544
15.3 Comparaison avec d'autres systèmes axiomatiques (Aleksander & Dunmall)	550
15.4 Postulats physicalistes	554
15.5 Comparaison avec l'information de Shannon	557
15.6 Ontologie : entre éléatisme et spinozisme	559
15.7 Chiasmes métaphysiques et principes épistémiques	562
16 Fondations d'une théorie méréologique de l'esprit	567
16.1 Critique de l'empirisme cognitivisme	568
16.1.1 Le Problème de la genèse de la conscience	570
16.1.2 Méthode d'investigation à double aspect	573
16.2 L'Expérience du cerveau isolé	575
16.2.1 Je pense donc Je lie	580
16.2.2 Le Soi et l'Inconscient	588
Conclusion	595
Le Mystérianisme perspectiviste	597
La Thèse forte et la Thèse faible	598
ChatGPT : une alchimie des savoirs	601
Apocalypse artificieuse ou Jugement dernier?	604
Bibliographie	609
Annexes	625
Annexe 1 : Une nouvelle écrite par ChatGPT	625
Consignes de la nouvelle	626
To Your Pens!	626
A Web of Unlikely Encounters	626

La Tisserande des murmures	628
Annexe 2 : Hypothèses détaillées du Théorème de Gödel	635
Annexe 3 : La Thèse de Church-Turing	637
Méthode effective	637
Théorème de Turing	639
Annexe 4 : Le paradoxe de Cantor	641
Comment définir la taille d'un ensemble infini ?	641
Approche 1 : l'inclusion	642
Approche 2 : la densité	643
Approche 3 : La Cardinalité	646
Propriétés contre-intuitives de l'infini	649
L'argument de la diagonale de Cantor	650
L'Ensemble des sous-ensembles	650
Généralisation à n'importe quel ensemble	652
Le paradoxe de Cantor	653
Annexe 5 : Les Nombres transfinis	657
Annexe 6 : La Théorie de l'information de Shannon	659
Point de vue historique et mathématique	660
Les approches de Hartley et de Wiener	660
L'approche de Shannon	662
Système axiomatique	663
Les mesures de Shannon	665
Entropie jointes et conditionnelles	665
Entropie relative	666
Information mutuelle	666
Annexes 7 : Essai - La Conscience est-elle morte ?	669
Présentation de l'essai	669

Introduction

*D'ici la fin du siècle, on pourra
parler de machines pensantes
sans s'attendre à être contredit.*

Alan Turing

Il ne se passe peut-être pas une année sans qu'une sommité du monde technologique, frappée tout à coup d'une idée foudroyante, ne défraie la chronique. « Eureka ! » écrit-il sur les réseaux sociaux, « La conscience fut ! » scande-t-il avant d'être licencié de Google¹. Ces Archimèdes de l'intelligence artificielle se targuent d'avoir donné naissance à une machine douée de pensées, de sensations, d'émotions, de créativité.

En dépit de l'enthousiasme invariable que soulèvent ces déclarations, les prétentions métaphysiques à propos de la machine s'apparentent trop souvent à un astucieux escamotage intellectuel. Il n'a d'ailleurs pas fallu attendre la venue de notre siècle, avec son cortège de chatbots à la sauce Terminator, pour voir naître de telles allégations. Peut-être vous souvenez-vous du Turc Mécanique, la machine construite par Johann Wolfgang von Kempelen en 1770 ? Cet automate fait de câbles et d'engrenages, à l'effigie d'un Turc enturbanné, défrayait les chroniques. On le voyait jouer aux échecs — et gagner le plus souvent — contre toutes sortes de personnages célèbres, en ce compris Napoléon Bonaparte, Catherine II de Russie ou Benjamin Franklin. Ce que la foule ignorait, c'est qu'un être humain de chair et d'os était caché à l'intérieur. Il actionnait les bras, les mains, les doigts de l'automate par un savant système de cordages.

En prenant appui sur cet exemple cocasse, il est légitime de se demander si les algorithmes d'aujourd'hui, auxquels on impute de l'imagination, de

1. Blake Lemoine a été licencié le 22 juillet 2022 à la suite d'allégations affirmant que le chatbot sur lequel il travaillait devait être considéré comme une personne à part entière. Il avait en effet rendu publique une conversation tenue avec LaMDA (modèle de langage pour les applications de conversation), ce qui constituait une violation des règles de confidentialité. Certains observateurs ont toutefois fait remarquer que Google avait désapprouvé les sous-entendus moraux impliqués par les affirmations de Blake. En 2020, en effet, Margaret Mitchell et Timnit Gebru avaient été licenciés après avoir voulu mettre en garde face aux dangers éthiques soulevés par l'IA (Courrier international, *Google suspend un ingénieur qui assure que son IA est « consciente »*, (23 juillet 2022).

la créativité ou un sens esthétique, ne cacheraient pas, eux-aussi, la main invisible d'une intelligence humaine ? Derrière les fresques romantiques de Dall-E ou Midjourney, les symphonies baroques d'Aiva et les poésies parnassiennes de ChatGPT-4 ne trouve-t-on pas la compilation statistique de millions de productions artistiques ?

Où que l'on tourne la tête, l'on est pris en étau entre l'admiration tout à fait légitime que suscite le chef-d'œuvre statistique révolutionnaire des IA ; et la déception face au manque apparent de compréhension réelle dont la machine se montre capable. Pour le dire autrement, un fossé semble se creuser entre les prouesses techniques de l'intelligence artificielle et notre compréhension, somme toute superficielle, du fonctionnement de la pensée humaine.

D'un point de vue philosophique, aborder la question de savoir si une machine peut penser est une façon commode de s'interroger sur ce que c'est que la « pensée », la « conscience » et « l'intelligence » ? Faute d'apporter une réponse à ces mystères fondamentaux — peut-être insolubles du reste — nos contemporains risquent de prendre pour acquis une double supposition : (1) celle d'après laquelle le cerveau fonctionnerait comme un ordinateur sophistiqué et (2) celle d'après laquelle la vie mentale serait produite entièrement par le cerveau biologique. Sous ces hypothèses, plausibles à première vue, mais pas nécessairement évidentes, toute évolution réalisée en informatique nous rapprochera d'une meilleure compréhension neurobiologique de l'esprit humain ; si bien que la méthodologie actuellement dominante en intelligence artificielle finira par éclipser les approches singulières de la neurobiologie et de la psychologie.

Le premier congrès sur l'intelligence artificielle, tenu à Dartmouth en 1956, conjecturait d'ailleurs que « tous les aspects de l'apprentissage et les autres dimensions de l'intelligence peuvent, en principe, être décrits de façon si précise qu'un ordinateur pourrait les simuler ». Il réunissait des chercheurs tels que Claude Shannon, Nathan Rochester, Ray Solomonoff, Allen Newell, Herbert Simon, Marvin Minsky ou John McCarthy. Le communiqué officiel présentait une sorte de manifeste qui prélude certainement à l'évolution actuelle de la discipline² :

On tentera de découvrir comment amener les machines à utiliser le langage, à former des abstractions et des concepts, à résoudre des types de problèmes actuellement réservés aux humains, et à s'améliorer elles-mêmes.

Cette idée d'après laquelle l'esprit humain fonctionne comme un ordinateur est ce que j'appelle la réduction *ad computatum* (la réduction au calcul). En l'état actuel, il s'agit d'une simple hypothèse méthodologique. Bien qu'elle semble trouver un certain écho auprès de la population générale,

2. La traduction est de moi. Le texte est issu de *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1955.

il semble toutefois difficile de donner des arguments définitifs. Je tenterai d'épingler, tout au long de cette thèse, un certain nombre de problèmes au-devant desquels cette posture philosophique, si elle n'est pas suffisamment interrogée, risque de nous entraîner.

Pour critiquer la réduction *ad computatum*, il importe tout d'abord de comprendre ce que les algorithmes contemporains sont capables de faire, de sorte à souligner, dans un second temps, quelles sont leurs faiblesses. Le point de départ de cet ouvrage consistera donc à reconnaître à l'intelligence artificielle deux capacités indéniables : celle d'analyser et celle de synthétiser des informations. Qu'est-ce que j'entends exactement par là ?

L'analyse, dans sa quintessence, consiste à raisonner de façon logique. En partant d'un certain nombre de prémisses, il s'agit d'effectuer des déductions et de se hisser, étape par étape, vers une conclusion. N'est-ce pas ce que réalisait déjà, en substance, la première machine à calculer de Blaise Pascal ? Si l'on s'en tient aux résultats, il y a indubitablement quelque chose, dans les algorithmes, qui s'apparente à — et surpasse même souvent — la réflexion humaine.

La *synthèse*, quant à elle, consiste à résumer de l'information et à en extraire quelque chose de nouveau. Elle comprend un double mouvement. Le premier, centrifuge, consiste à simplifier, à réduire, à abréger. Le second mouvement, centripète, consiste à fusionner, à réunir, à hybrider. N'est-ce pas ce que fait ChatGPT lorsqu'il résume le contenu d'un texte ? Ne serait-il pas foncièrement malhonnête, intellectuellement, que de refuser à nos réseaux de neurones la reconnaissance d'une certaine forme de créativité, d'arborescence et de génération ?

La question résiduelle est celle de savoir si les paradigmes actuels d'intelligence artificielle, au-delà de leurs capacités d'analyse et de synthèse, sont en mesure de *juger*. Le jugement est, en effet, une aptitude dont l'être humain se montre généralement capable. C'est la capacité à généraliser et à abstraire des informations de façon à distinguer le vrai du faux, à délibérer et à prendre une décision bien fondée. C'est le point de charnière entre l'analyse et la synthèse, entre le raisonnement et l'expérience.

Au sein d'une analyse, le *jugement* consiste à déterminer si les prémisses de départ sont vraies et si le résultat du calcul fait sens. Il s'agit donc de donner une signification aux termes du discours, de façon à ne pas perdre pied avec la réalité désignée par les variables du calcul. C'est ce que vous faites lorsque, après avoir résolu une équation, et représenté chaque variable du problème par des symboles alphanumériques, vous vous demandez ensuite si les valeurs obtenues ont un sens. Vous contextualisez votre raisonnement, vous interprétez la solution et vous la critiquez. Est-ce une bonne approximation ? Une de mes hypothèses était-elle fausse ? Les principes physiques utilisés s'appliquent-ils au contexte concerné ? Le jugement est donc ce qui permet de manipuler des abstractions sans perdre de vue la réalité désignée par cette abstraction.

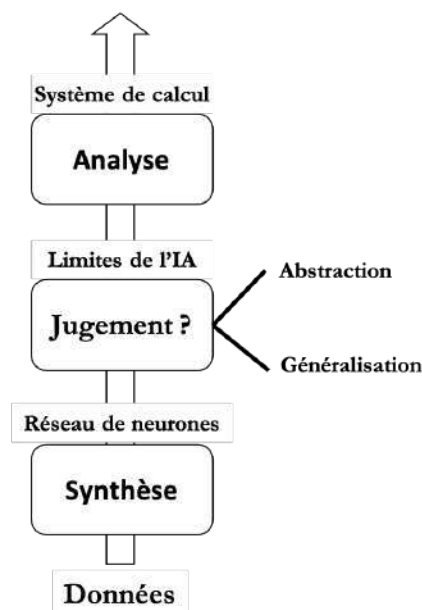


FIGURE 1 – Le jugement se présente comme le niveau intermédiaire entre l'analyse et la synthèse.

Au sein de la synthèse, le *jugement* consiste à ramener la diversité des informations à des concepts. C'est donc une capacité de généralisation. Vous y recourez notamment lorsque, tentant de décrire le style pictural d'un artiste, telle que Frida Khalo, vous parlez « d'art figuratif », « d'expression de la souffrance » ou de « réalisme magique ». Ces expressions ont un sens pour vous et (espérons-le) pour vos interlocuteurs. Vous seriez toutefois bien en peine d'expliquer exactement ce que ces concepts désignent, à l'aide de mots et de définitions exactes. Vous pouvez certes donner des éclaircissements à ce que vous entendez par « réalisme magique », mais pas une définition ayant la clarté irréfutable d'une formule mathématique. Il y a deux raisons à cela. La première, c'est que votre mémoire n'est pas assez vaste pour embrasser, d'un seul coup, toutes les peintures figuratives qui échoient au réalisme magique. La seconde raison, c'est qu'il se pourrait que le concept de « réalisme magique » ne soit pas la simple collection nominaliste des peintures recevant cette appellation, mais qu'il désigne quelque chose de plus. Une possibilité ? Une structure ? Un ensemble de propriétés ? C'est ce qu'il faudra interroger.

La question soulevée par cette thèse peut recevoir la formulation suivante :

L'intelligence artificielle possède-t-elle cette capacité de jugement qui semble tenir l'exact milieu entre la faculté de synthèse et la faculté d'analyse ? Dans le cas contraire, peut-elle comprendre et manipuler des concepts tels que le « réalisme magique » ?

Motivations éthiques

La question de savoir si une machine possède la faculté de juger semble d'autant plus importante, aujourd'hui, que de plus en plus de décisions sont confiées aux algorithmes. Lorsqu'une banque doit faire un prêt à un client, qu'un établissement scolaire doit évaluer les capacités d'un enseignant, qu'un juge doit estimer si le condamné est susceptible de récidiver, il est possible de se référer à l'évaluation d'un programme. Celui-ci utilise les données concernant la personne (son nom, son adresse, son casier judiciaire, sa profession, etc.) afin de déterminer s'il s'apparente statistiquement à des profils ayant des bonnes performances. Au plus la base de données de l'algorithme est riche en information, au plus précise sera l'évaluation rendue.

Or les données que nous utilisons ne sont pas forcément justes et équitables. Aux États-Unis, les personnes Afro-Américaines sont souvent jugées plus sévèrement que les personnes blanches. Face à un même crime, des sanctions plus élevées seront observées. Il s'ensuit que si l'on nourrit l'algorithme en lui soumettant les données réelles, il risque de reproduire une plus grande sévérité à l'égard de ces personnes. Même si les variables sensibles, telles que l'ethnicité ou la couleur de peau, ne sont pas injectées explicitement dans la base de données, des corrélations peuvent être établies à partir de données non sensibles. L'adresse postale, le nom et le prénom, à titre d'exemple, sont souvent corrélés à l'ethnicité. Le programme peut déterminer que les personnes du nom de Medhi et ressortissants de la banlieue sont plus susceptibles de commettre des crimes.

Cette discrimination n'est pas facile à éviter avec les algorithmes que nous utilisons aujourd'hui, et ce pour trois raisons. La première vient de ce que les sociétés à la pointe de la technologie ne sont pas particulièrement incitées, financièrement et légalement, à investir massivement dans des algorithmes éthiques libres de tous biais algorithmiques. La seconde difficulté vient des données elles-mêmes. Les big data sont si massives qu'un être humain n'a pas assez, d'une vie humaine, pour consulter tout ce qui est dit sur Instagram en une journée. La troisième raison, pour finir, est intrinsèque aux méthodes de programmation les plus en usage : les réseaux de neurones. Il n'est pas facile, au sein de ces réseaux, d'éviter que ne soient produites des corrélations indésirables entre les variables sensibles et les variables non sensibles.

J'aimerais illustrer le problème des biais algorithmiques en étudiant les réponses produites par un algorithme d'assistance juridique américain nommé COMPAS (Correctional Offender Management Profiling Alternative Sanctions). Cet outil, introduit dans les tribunaux en 2016, est loin d'être un juge impartial. Sur l'image ci-contre, vous constaterez l'évaluation du « risque de récidive » rendu par l'algorithme à l'égard de deux personnes. Les deux prévenus, Vernon Prater et Brisha Boden, ont été arrêtés pour des motifs similaires : le premier a voulu voler des outils dans un magasin. La



FIGURE 2 – Algorithme assignant un risque de récidive chiffré à deux prévenus (2014).

seconde, craignant d'être en retard pour chercher sa belle-sœur à l'école, a voulu « emprunter » un vélo et un scooter sans propriétaire le long du trottoir, avant de finalement changer d'avis. Pas de chance pour elle : un voisin visiblement soucieux de faire respecter l'ordre public — mais habilement dissimulé derrière les tentures de son salon — l'a dénoncée à la police !

Dans les deux cas de figure, le larcin est estimé à hauteur de 80 dollars. Cependant, Vernon est un homme blanc de quarante ans qui a déjà été arrêté précédemment pour deux vols à main armée. Brisha est une femme noire de dix-huit ayant, pour seuls antécédents judiciaires, des infractions juvéniles. Au moment de calculer le risque de récidive, le système a donné un 3 à Vernon et un 8 à Brisha. Plusieurs années plus tard, Vernon s'est fait arrêter pour une attaque à main armée, tandis que Bodon n'a pas commis d'autre crime. Il est plausible de penser que la plupart des êtres humains, exception faite des membres du Ku Klux Klan, n'auraient pas commis cette erreur. Cela ne veut évidemment pas dire que l'algorithme soit raciste, au sens où il discrimine nécessairement les personnes Afro-Américaines, mais ça illustre que les réponses statistiques peuvent suivre une logique qui ne coïncide pas avec la jurisprudence en vigueur³.

Les algorithmes ont foncièrement évolué depuis 2014, mais ils sont toujours sujets à des méprises. Vous souvenez-vous de cette application sur smartphone, sortie en 2021, qui disait d'une femme noire que c'était un gorille ? Plutôt que d'être corrigée, l'option consistant à reconnaître des singes a été retirée de l'application. Il n'est pas toujours évident de déterminer, à

3. Notons qu'il y a également un manque de transparence concernant la façon dont l'algorithme COMPAS est utilisé, ou non, dans les décisions qui sont rendues. Ce cas est bien documenté dans un article de ProPublica (<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>). Pour une analyse plus poussée des biais algorithmiques, se référer à l'ouvrage de O'NEIL intitulé *Algorithmes : la bombe à retardement*. Pour une analyse plus poussée des black boxes, se référer à *The age of secrecy and unfairness in recidivism prediction* (Cynthia Rudin, Caroline Wang, Beau Coker, 2018).

coup sûr, l'origine d'une erreur de façon à l'enrayer. Les réseaux de neurones utilisés aujourd'hui fonctionnent d'une manière telle qu'il est hélas difficile de fournir des garanties. Même lorsque le système présente de bons résultats sur sa base de données de test, il n'est pas exclu qu'une erreur soit commise en présence d'un cas réel. En dépit de ce défaut, il reste commode de donner des décisions aux algorithmes, et ce pour deux raisons.

D'une part, les programmes informatiques bénéficient d'une sorte d'aura d'objectivité. L'opinion est fermement enracinée que la machine, peu sujette aux sautes d'humeurs ou aux traitements de faveur, opère partout avec la même impartialité.

D'autre part, confier la décision à une machine permet de se dédouaner de certaines responsabilités. Il est plus aisé, en cas de contestation, d'incriminer la machine qu'un être humain de chair et d'os. Du reste, la loi ne prévoit pas toujours clairement qui doit être tenu pour coupable en cas de défaillance algorithmique : son utilisateur ? le concepteur du programme ? la firme utilisant la machine ? ceux et celles qui ont produit ou introduit les données ?

Il semble donc capital de saisir comment la machine pense, et comment elle décide de classer certaines personnes dans telle ou telle catégorie. Si j'arrive à montrer qu'il y a quelque chose, dans le jugement humain, qui ne peut pas être imité par les algorithmes, je détiendrai là un argument appuyant le fait que les décisions confiées aux machines doivent toujours être *supervisées* par un être humain.

Proposition 1 (Principe de supervision). *Les décisions algorithmiques peuvent assister le jugement humain, mais non s'y substituer. Lorsqu'une tâche est suffisamment simple pour envisager son automatisation complète, le protocole doit néanmoins prévoir la possibilité d'une intervention humaine.*

Un problème philosophique

L'un des problèmes fondamentaux de l'intelligence artificielle consiste à déterminer comment les données à disposition de la machine (par exemple l'ensemble des données concernant un repris de justice) peuvent être catégorisées de façon à former des concepts manipulables (coupable, non coupable, suspicieux). Cette capacité de réunir la diversité des données de l'expérience sous une seule catégorie manipulable par les lois de la pensée est ce qu'on appelle, en philosophie, le *jugement*.

Ce problème que je viens de décrire ne relève pas d'une simple question technique, mais s'appuie sur une énigme épistémologique : celle qui consiste à se demander quel rapport existe entre les choses sensibles et les concepts que nous en avons. C'est une chose en effet que de voir un criminel, c'est tout à fait autre chose de dire qu'il s'agit d'un « criminel ». Dans le concept de « criminel » sont contenus tous les crimes possibles et imaginables, aussi

bien celui que l'on a sous les yeux que ceux qui se cachent, sournoisement, dans les tréfonds de la nature d'un criminel. En affirmant une phrase aussi banale que : « c'est un criminel ! », je distingue un être humain particulier du reste de la foule pour lui donner une unité, une forme individuelle. Dans le même temps, je l'assimile aussi, quant à ses ressemblances, à tous les autres criminels de l'histoire de l'humanité. J'isole au sein de ma perception de cet homme l'existence d'un visage torturé, d'un regard fuyant, d'un passé difficile, d'un manque d'empathie, d'une faiblesse de caractère, d'une perversion dissimulée, d'un esprit du mal. Il s'établit dès lors une relation d'association et de différenciation entre cet instant de perception singulier (j'expérimente quelque chose dans l'ici et maintenant) et une connaissance universelle (je dis de cette chose expérimentée que c'est un visage, une expression, une voix de criminel).

Ce problème que j'évoque est connu, en philosophie, sous le nom du « problème des universaux ». Il consiste à se demander pourquoi l'esprit est capable de penser le rapport entre un concept universel et un objet singulier, et d'après quelles lois ce rapport peut se produire. De ce problème fondamental découlent de nombreuses interrogations : les concepts que nous manipulons ont-ils une réalité à part entière ou ne sont-ils que des expressions commodes, des raccourcis linguistiques, pour désigner les objets du monde ? Aurait-on la capacité de voir un objet tel qu'un arbre si l'on n'en avait pas le concept ? Détiendrait-on le concept de cet arbre si nous n'avions jamais été exposés à la perception d'arbres particuliers ?

En termes informatiques, ces questions relatives aux universaux consistent à se demander s'il existe quelque chose telle que des données « brutes », qui peuvent être collectées sans prescrire de structures prédéfinies ? Ou s'il est au contraire indispensable d'avoir des catégories préalables pour représenter les données ? Que par conséquent il n'existe pas de données « brutes », mais que toutes les données sont « mises en forme » ?

Deux réponses sont traditionnellement envisagées au problème des universaux :

- **Réponse empirique :** Si l'on défend que ce sont les objets qui sont premiers et que les concepts ne sont obtenus que par l'expérience directe du monde, l'on se range dans la tradition *empirique*. L'esprit humain n'est alors rien que le résultat d'un frottement, d'une confrontation, entre le monde réel et notre représentation de celui-ci. Si nous avons le concept d'arbre, de loup, de maison et de criminel, et non un concept pour distinguer aisément les différentes odeurs, c'est parce que ces concepts-là se sont avérés les plus adaptés au développement de l'espèce humaine dans les conditions qui étaient celles du Paléolithique.
- **Réponse idéaliste :** Si l'on défend au contraire que ce sont les

concepts qui nous permettent d’appréhender le monde ; que sans ceux-ci nous marcherions à tâtons dans la nuit ; que le monde sensible, loin d’être peuplé d’objets individuels distinguables, ne serait plus rien qu’un patchwork, un brouhaha digne de la soupe cosmique primitive des premiers âges de l’univers — alors on se range dans la tradition *idéaliste*. Celle-ci stipule qu’il existe des structures *a priori* de la pensée, des formes pures, qui donnent une intelligibilité à notre expérience subjective. Les concepts, pour le dire autrement, sont les lunettes indispensables à la perception, clair-obscur, du monde réel.

Dans le cadre de cette thèse, je ferai l’hypothèse que la tension existant entre l’empirisme et l’idéalisme, loin d’être reléguée aux activités oiseuses — et déjà fort nombreuses — de la philosophie, a donné naissance à deux paradigmes de programmation en intelligence artificielle : la programmation top-down et la programmation bottom-up. La première part des concepts comme s’il s’agissait d’une réalité préexistante et tente de dire des choses sur le monde réel. La seconde part au contraire des objets et adapte les concepts au gré de l’expérience^[4].

Je vais détailler ces deux paradigmes dans les deux prochaines sections. Mais auparavant, j’aimerais clarifier, en une formule synthétique, l’hypothèse qui est au fondement de ma démarche :

Hypothèse 1. *La question de savoir comment se forment les concepts est le nœud gordien de l’intelligence artificielle. Le point de vue empirique stipule que les concepts s’adaptent au monde qui nous entoure et évoluent au gré de l’expérience. Le point de vue idéaliste stipule que ce sont les objets du monde qui se conforment à nos concepts. Ces deux points de vue ont donné naissance aux paradigmes de programmation top-down et bottom-up, dont il va être question dans les sections suivantes.*

Programmation top-down

Lorsque l’on adopte le point de vue idéaliste, on suppose que les concepts ont une existence objective. Ils gouvernent, ils filtrent, ils orientent notre représentation du monde. Il s’ensuit que si l’on cherche à programmer un algorithme visant à exécuter une certaine tâche, on commencera par définir

4. L’usage que je fais des termes « programmation top down » et « programmation bottom up » est quelque peu différent de celui qu’on retrouve parfois dans la littérature. *The Perl Design Patterns Book*, utilise cette terminologie pour désigner la façon de construire l’architecture d’un programme écrit à partir de règles, en distinguant les cas où ces règles s’appliquent à des situations particulières ou à des situations générales. J’ai toutefois pris la décision de redéfinir quelque peu ces termes de façon à ce que toutes les approches sous forme de règles prescriptives soient comprises dans l’appellation top-down, tandis que le mot bottom-up s’applique aux méthodes de Deep Learning s’appuyant notamment sur les statistiques et l’auto-apprentissage. D’après ma définition, la grande majorité des programmes des années 80 sont donc top-down, tandis que les réseaux de neurones, par exemple, sont bottom-up.

les concepts liés à cette tâche, après quoi seulement on confrontera la machine à des cas particuliers. Dans cette approche de programmation, il est fait l'hypothèse implicite que l'analyse logique des concepts suffit à juger du vrai et du faux dans les contextes réels.

J'aimerais illustrer ce paradigme à l'aide d'un exemple issu de la science-fiction. Supposons que l'on veuille façonner un robot moral similaire à ceux auxquels nous a habitué la plume d'Isaac Asimov. On ne cherchera pas à construire les concepts de « bien », « d'équité » ou de « justice » en partant de la pratique sociale, ou en imitant simplement les comportements humains moyens. On partira au contraire du principe que la morale est une réalité objective : elle est dotée de lois auxquelles la société doit se conformer. Dès qu'on aura identifié ces lois, il suffira de les écrire dans un langage informatique. Chaque terme de la loi morale doit pouvoir être exprimé par une suite de nombres binaires. Chaque décision, pour sa part, doit être présentée comme le résultat d'un calcul opéré sur ces nombres.

L'algorithme, muni de cette nouvelle grammaire, peut manipuler les règles morales pour déduire, *a priori*, quelles sont les réponses moralement admissibles face à un dilemme donné. Tout se passe, pour lui, comme si le bien se ramenait à des preuves mathématiques. La générosité, la compassion, la justice sont tout autant de « théorèmes de la morale » susceptibles de preuves. Ce genre d'algorithme qui analyse, déduit et raisonne d'après les strictes lois de la logique, est communément appelé un Système Expert. On trouve les premières traces de ces Systèmes Experts dans les années cinquante, aussi bien en ce qui concerne la discipline du jeu d'échecs que la médecine, sans oublier la consultance financière. Ce sont eux encore qui inspirent à Isaac Asimov les lois morales de la robotique, reprises dans le film *I Robot* avec Will Smith dans le rôle principal, et formulées ci-dessous :

1. Un robot ne peut porter atteinte à un être humain ni, restant passif, laisser cet être humain exposé au danger ;
2. Un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres entrent en contradiction avec la première loi ;
3. Un robot doit protéger son existence dans la mesure où cette protection n'entre pas en contradiction avec la première ou la deuxième loi.

Ces lois sont pensées de façon tout à fait abstraite et universelle. Le programmeur n'a pas eu besoin de se référer à un contexte particulier pour les établir. Il laisse au robot le soin d'identifier, au sein de chaque contexte, comment les appliquer pour le mieux. Cela demande de faire preuve d'interprétation, de pondération et, in fine, de jugement. Or, ce sont précisément ces aptitudes dont on se demande si elles peuvent être encodées sous forme de lois. L'analyse logique suffit-elle à saisir un contexte social et à comprendre les implications morales ? Ou un robot programmé de cette manière

sera-t-il une sorte de caricature autistique, une marionnette loufoque, un fieffé logicien incapable de comprendre le langage non verbal ?

Le paradigme de programmation qui part de lois analytiques et qui suppose que le contexte doit s'y soumettre est ce que j'appelle le paradigme *top-down*. L'avantage de la programmation top-down, c'est que l'algorithme fonctionne de façon transparente. Les lois régissant le modèle étant parfaitement connues, il est permis de remonter, étape par étape, au principe qui est à l'origine du comportement de la machine. Si un robot refuse d'obéir à un ordre, on saura que cet ordre, selon ses analyses, met en danger la vie d'un être humain. C'est pourquoi l'on admet généralement de ces algorithmes qu'ils sont traçables et justifiables.

Cette traçabilité dont je parle ne veut pas dire que tout ce que fait la machine est absolument prévisible. Bien que les lois soient connues par avance, on ne peut pas toujours déterminer, par avance, à quel résultat cela aboutira. Cette imprévisibilité logique est comparable à ce que l'on appelle, en mathématique, les comportements chaotiques. Deux jeux d'équation identiques, confrontés à des contextes similaires, peuvent donner naissance à des réponses totalement différentes. Quiconque s'est piqué, un jour dans sa vie, de prédire la météo ou les cours boursiers en aura sûrement fait les frais. Le plus petit facteur peut modifier sensiblement le comportement de l'ensemble.

Asimov semble conscient de ce phénomène. Dans sa nouvelle intitulée *Les Robots et l'Empire*, l'androïde R. Daneel Olivaw déduit, sur base des lois initiales de la robotique, l'existence d'une loi zéro consistant à ne jamais nuire à l'humanité. Il décide de violer les lois précédentes en faveur de ce principe supérieur. Ce faisant, il ne fait rien d'autre qu'appliquer strictement le code déontologique qui lui a été enseigné. Son comportement est inattendu, voire chaotique, mais il est complètement traçable et justifiable — ou tout au moins justifié.

Des surprises analogues peuvent survenir dans des applications bien différentes du Robot moral. En s'appuyant sur la méthode top-down, il est par exemple permis de dériver des démonstrations mathématiques qui, quoique tout à fait logiques et prédictibles, ne seraient pas forcément venues à l'esprit de l'être humain. Un programme appelé le *Logic Theorist*, développé par Herbert Simon et Allen Newell, s'est montré capable de prouver certains théorèmes des *Principia Mathematica* — un ouvrage de logique mathématique réputé pour sa haute technicité. Le programme donnait même parfois des preuves plus élégantes que celles connues jusqu'à présent⁵. Au lieu de prendre pour fondement de son raisonnement les lois d'Asimov, le système s'appuie sur les axiomes de l'arithmétique. Il part du résultat à démontrer et applique des modifications jusqu'à se ramener à un axiome admis.

5. Ce résultat a en fait été obtenu par le Logic Theorist, l'ancêtre du General Problem Solver. Ce logiciel a produit une preuve du théorème 2.85 des Principia Mathematica, jugée plus courte et plus simple que celle de Whitehead et Russell.

Ce gain de transparence et de compréhension dont je viens de parler se paye néanmoins par un manque d'adaptabilité. Le programme top-down, qu'il soit spécialisé à résoudre des dilemmes moraux ou des problèmes mathématiques, est en effet incapable de former des concepts nouveaux. Il ne peut pas non plus penser en dehors des contextes couverts par ses lois analytiques. Que fera le robot d'Asimov si la vie d'un animal est en danger ? Ou si l'enjeu éthique concerne la sauvegarde de l'environnement ? La machine top-down est un cheval avec des œillères. Elle suit les ornières qui lui ont été attribuées, sans faire de pas de travers.

C'est pourquoi il peut sembler judicieux de se demander, à rebours de l'approche idéaliste, ce que devient la programmation lorsque l'on adopte le point de vue d'après lequel les concepts n'ont pas d'objectivité ; que ce sont des idées qui se forment au contact de l'expérience ?

Programmation bottom-up

On a vu que les robots d'Isaac Asimov sont programmés pour suivre une loi morale prédéfinie. Ils possèdent des concepts universaux et rationnels qu'ils tentent d'appliquer au monde réel. Cette approche est compatible avec le point de vue idéaliste. L'on est toutefois en droit de se demander ce que deviendraient ces robots si l'on adoptait une perspective empirique ? Dans ce cas de figure, en effet, le bien moral n'aurait pas d'existence objective. Faute de pouvoir se raccrocher à une norme écrite dans le marbre de toute éternité, c'est en partant de l'observation des comportements humains qu'il sera possible de déterminer ce qui est juste ou non.

À quels types de stratégie cette vision empirique peut-elle bien conduire ? S'il était né aujourd'hui, Asimov se serait sans doute moqué des lois de la robotique. Il les aurait jugées trop étroites, trop inflexibles. Plutôt que de programmer des règles analytiques, il se serait contenté de fournir à la machine le plus grand nombre possible de données à synthétiser. Il aurait fait le pari que le robot, après un entraînement adapté, est en mesure d'identifier par lui-même les comportements qui doivent être tenus pour « justes » et ceux qui doivent être vus comme « iniques ». Pour booster l'apprentissage artificiel, il aurait sans doute eu recours à la bonne vieille méthode du bâton et à la carotte. Les bonnes réponses sont encouragées, les mauvaises sont éliminées. À chaque nouvelle génération de robot, on s'attend à voir les performances morales croître aussi sûrement qu'un rosier exposé à la faveur du soleil printanier.

Cette façon de faire est fort utilisée dans la conception de chatbots. Ces derniers sont capables de s'exprimer à propos des choses morales ou de tout autre sujet auquel ils auront préalablement été exposés. Le premier agent conversationnel un tant soit peu convainquant est issu des années septante. Ce programme, nommé Eliza, recevait des phrases à analyser et tentait d'y

reconnaître des formes grammaticales : sujet, verbe, complément, etc. Il était conçu pour associer, à chaque phrase qu'on lui soumettait, une réponse statistiquement cohérente. Ces associations, on s'en doute, n'avaient pas la subtilité de ChatGPT-4. Le programme parvenait tout de même, non sans ironie, à se faire passer pour une caricature du thérapeute rogérien — connu pour rephraser les propos du patient de façon pseudo-intelligente. Voici un extrait d'une conversation que j'ai eue avec Eliza :⁶ :

ELIZA : Comment allez-vous ? S'il vous plaît, dites-moi votre problème.

VOUS : J'essaie de vous décrire pour un livre que j'écris.

ELIZA : Aimez-vous écrire ?

VOUS : Oui, mais pas nécessairement à votre propos.

ELIZA : C'est assez intéressant.

VOUS : Que puis-je écrire sur vous ?

ELIZA : Peut-être que, dans vos fantasmes, nous écrivons l'un sur l'autre.

Les discussions avec Eliza n'ont naturellement jamais passé le test de Turing. Il suffit de quelques minutes seulement pour qu'un interlocuteur moyen réalise qu'il n'a pas face à lui un être humain de chair et d'os, mais une machine de cuivre et de silicium. Les méthodes statistiques derrière les chatbots des années 60 vont toutefois jouir de plusieurs innovations significatives qui conduiront, en 2022, à la naissance du ChatGPT que nous utilisons aujourd'hui. Ces innovations sont les réseaux de neurones (introduits dès les années 50 et perfectionnés dans les années 80) et les transformeurs (en 2017).

Les Réseaux de neurones

Les réseaux de neurones ont été mis à l'honneur par le Prix Nobel de physique 2024, décerné à John Hopfield et Geoffrey Hinton pour leurs travaux pionniers. Leur approche consiste à traiter l'information au moyen de régressions statistiques élaborées. Cette régression vise à simplifier l'information reçue en observant des tendances. Si, par exemple, on vous donne un nuage de points correspondant au nombre d'actes criminels commis par un individu en fonction de son revenu, vous pouvez tenter de tracer une droite qui dira : en-dessous d'un certain revenu, il y a au moins deux crimes qui sont commis en moyenne sur la durée d'une vie. Cette méthode, assez simple, n'est jamais qu'une approximation statistique. On est encore loin de la fiction imaginée dans le film *Minority Report*, où des êtres dotés de pouvoirs spéciaux (appelés précogs) sont capables de prédire l'occurrence d'un crime avec une précision absolue.

Il va sans dire que les situations réelles ne présentent pas qu'un seul axe d'analyse, celui du revenu, mais doivent tenir compte de toutes les autres

6. J'ai construit cet exemple en me rendant sur une page reprenant le code d'Eliza : <https://www.masswerk.at/elizabot/eliza.html>

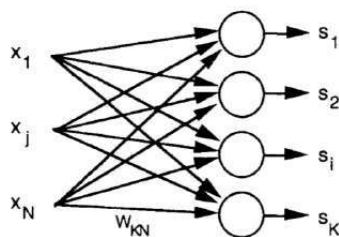


FIGURE 3 – Le réseau de neurones connecte les neurones initiaux à gauche (les données du problème) aux neurones terminaux à droite (les réponses admissibles).

données relatives au casier judiciaire du prévenu : le nombre d’actes criminels déjà commis, la nature de ceux-ci, la durée d’incarcération, etc. Ce même procédé est utilisé par les réseaux de neurones. Chaque donnée du problème est représentée par un neurone initial : il existe donc un neurone représentant le revenu de la personne, un autre représentant son âge, un autre correspondant au nombre de crimes déjà commis, etc. Ces neurones initiaux sont connectés les uns les autres. Ils sont également connectés à des neurones terminaux, qui correspondent aux réponses permises : « le prévenu va commettre un crime » ou « il ne va pas commettre un crime ». Il peut aussi exister un neurone terminal pour chaque sous-catégorie de réponse : « criminel probable », « criminel improbable », « criminel très improbable », etc. Le système est entraîné de façon à dire quel neurone terminal doit s’activer en fonction de l’information dont sont porteurs les neurones initiaux. Cet entraînement est purement statistique. Le réseau crée des liaisons qui épousent le plus fidèlement possible les réponses qui ont été observées dans les contextes connus. Il espère ainsi être en mesure d’apporter des réponses fiables face aux situations nouvelles.

Le principal défi technique soulevé par les réseaux de neurones consiste à lier adéquatement les informations. Tant que les mots d’une phrase ou les données relatives à un individu criminel sont regardés de façon isolée, il est difficile de comprendre le sens de la phrase ou de déterminer la culpabilité du prévenu. Pour affiner le jugement de la machine, et s’approcher de ce que fait l’être humain lorsqu’il lit un document ou qu’il analyse le dossier judiciaire de l’accusé, il est crucial de parvenir à lier les termes du problème et à modéliser la façon dont ils s’influencent les uns les autres.

Les Transformeurs

C’est justement pour répondre à ce problème de l’articulation des informations neuronales que s’est développée la deuxième technologie phare, celle des transformeurs. Dans un article d’Ashish Vaswani, publié en 2017, et intitulé « L’attention est tout ce dont vous avez besoin », un modèle de réseau

de neurones est présenté, ayant pour tâche d'analyser le langage. Sa particularité consiste à ne pas traiter la phrase à analyser de façon séquentielle, mais de façon parallèle. Au lieu de regarder les mots lettre par lettre et les phrases mot par mot, l'algorithme survole la phrase en s'inspirant des processus attentionnels de l'être humain.

La grande idée des transformeurs est de déterminer en quoi la présence d'un mot au sein de la phrase influence la compréhension des mots adjacents. Il va sans dire que l'algorithme, comme la plupart des réseaux de neurones, ne fonctionne que si on lui donne une très grande base de données : il peut s'agir des pages Wikipédia, des sites Web répertoriés sur Google ou des messages d'utilisateurs prélevés avec des applications sur smartphone. Le programme va utiliser tous ces exemples de phrases connues pour affiner la façon d'orienter son attention au sein de la phrase. Pour le résumer de façon lapidaire, on pourrait dire que le transformeur est un prédicteur qui consiste à déterminer, au sein d'un bout de phrase, quel mot doit suivre. Il s'apparente à ce titre aux exercices de grammaire :

Pablo est ... dans la classe.

Marc n'a pas ... à son ordre.

L'approche du transformeur, fondée sur l'association statistique et l'apprentissage, s'inscrit dans ce que j'appelle la programmation *bottom-up*. L'avantage de la programmation bottom-up, c'est la grande diversité des réponses observées. La machine est libre de découvrir de nouvelles relations. Le grand défaut, c'est un manque de contrôle et de transparence. Si vous donnez à un chatbot les conversations issues des réseaux sociaux, il y a des risques que l'algorithme finisse par devenir le fer de lance des fake news. C'est d'ailleurs ce qui a été observé avec les toutes premières versions de ChatGPT, qui tenaient des propos discriminatoires. Certains se souviendront peut-être aussi du scandale de BlenderBot 3 ? Ce chatbot, lancé le 5 août 2022 par Meta, affirmait mordicus que Donald Trump était toujours président des États-Unis... et qu'il y avait beaucoup de preuves comme quoi les élections avaient été truquées. Quoique factuellement fausses, force est de constater que ces allégations avaient quelque chose de tristement prophétique.

Le Jugement intuitif

Il semblerait que le paradigme top-down et le paradigme bottom-up se tiennent en vis-à-vis. Ils sont comme le ciel et la terre, comme la lune et le soleil, comme la pluie et l'éclaircie, l'averse et la tempête. L'approche *top-down*, en effet, part des lois connues de la théorie, encodées par avance,

et cherche à en dériver des conclusions utiles. La validité de ses réponses ne dépend cependant de rien d'autre que de la validité des prémisses initiales. Lorsque l'on vous demande « quelle est la couleur du cheval blanc de Napoléon ? », la réponse se trouve déjà inscrite dans la question. De la même manière, lorsque la machine doit évaluer une position de jeu d'échecs pour déterminer s'il est possible « de faire mat en cinq coups », la réponse se trouve inscrite, de toute éternité, dans les lois de la logique. On dit que la machine produit des jugements analytiques *a priori*. *Analytiques*, car ils découlent de l'analyse logique des termes du problème. *A priori*, car il n'est pas nécessaire de se confronter à l'expérience du monde.

L'approche *bottom-up*, en revanche, part de l'expérience sensible. Elle brasse de grands ensembles de données pour établir des corrélations. Elle tente donc de chercher à faire des rapprochements qui ne sont pas contenus, *stricto sensu*, dans la définition des termes. Si elle observe que les généraux sont généralement juchés sur des chevaux blancs, elle peut en déduire que Napoléon l'était aussi. Si elle observe que certaines configurations du jeu d'échecs mènent généralement à la défaite des pièces blanches, elle va décréter que la position est perdante pour les blancs, sans avoir calculé combien de coups seront nécessaires pour obtenir échec et mat. Ce raisonnement statistique n'est pas une vérité logique. Par suite de quoi, les jugements qui en résultent ne sont pas analytiques *a priori*, mais ils sont synthétiques *a posteriori*. *Synthétiques*, car on arrive à une vérité qui dépasse la définition des termes du problème. *A posteriori*, car il est nécessaire, pour établir un tel jugement, de se référer aux données de l'expérience concrète.

Cette différence que j'évoque entre les jugements analytiques et les jugements synthétiques est analogue à la différence existant entre l'induction et la déduction. En philosophie, l'induction désigne le fait de généraliser une observation récurrente pour en faire une loi générale. Il s'agit donc d'une espèce de *synthèse* de l'expérience. Elle se produit lorsque, après avoir contemplé les prairies écossaises par la fenêtre d'un train, vous affirmez à votre voisin de wagon que « Dans ce pays, tous les moutons sont blancs ». Il se peut tout à fait que vous ayez raison de résumer ainsi votre observation des paysages champêtres, mais il serait plus prudent de dire « Tous les moutons que j'ai vu à travers la fenêtre sont blancs ». Car si le profilé d'un mouton noir venait à être observé par votre voisin, posté à la fenêtre opposée, vous seriez obligé de réviser votre jugement et de dire, non sans une once d'ironie toute britannique : « Dans ce pays, il existe au moins un mouton dont une face n'est pas blanche. »

Au contraire de l'induction, la déduction est un raisonnement fondé, non sur l'observation de récurrences, mais sur les lois de la logique. Il s'agit donc d'une vérité certaine, comparable en cela au raisonnement *analytique*. Pour vous en donner un exemple simple, l'induction est ce à quoi recourt l'inspecteur Rouletabille lorsque, constatant qu'un crime a été commis dans la chambre jaune, et en admettant qu'il n'existe aucune sortie à cette chambre,

il en conclut que le criminel ne peut pas faire partie des personnes qui étaient dehors au moment des faits. La justification de ce raisonnement vient de ce que nul ne peut être à la fois *dans* la pièce et *en dehors* de la pièce. Admettre que le criminel ait le don d'ubiquité ne serait pas seulement contraire à l'expérience du monde ; cela irait aussi à l'encontre du fameux principe de contradiction, qui stipule qu'on ne peut pas avoir, en même temps, une chose et son contraire.

Les déductions et les inductions sont, toutes les deux, à la portée des processus de nos algorithmes. Tant que l'on s'en tient à des raisonnements logiques, en effet, l'être humain n'est pas en mesure de rivaliser avec les variantes, plus ou moins sophistiquées, de cette machine à calculer qu'on appelle le Système Expert. De la même manière, tant qu'on s'en tient à établir des corrélations statistiques, il est difficile de rivaliser avec la haute associativité des réseaux de neurones. La question qui se pose, dès lors, consiste à savoir s'il existe un autre type de faculté que l'induction et la déduction ? Une faculté qui se tiendrait, pour ainsi dire, entre les jugements purement *analytiques* et les jugements purement *synthétiques* ?

Dans cette thèse, j'aimerais discuter la présence de deux aptitudes de l'être humain qui résistent encore à nos algorithmes : ces aptitudes sont ce que j'appelle l'abstraction et la généralisation. En ce qui concerne *l'abstraction*, tout d'abord, on peut se demander comment la machine est en mesure de faire la liaison entre les concepts rationnels, définis de façon logique, et les catégories statistiques obtenues à partir de l'analyse des données. S'il voit de nombreux carrés, un réseau de neurones est certes capable d'identifier de nouveaux carrés, mais peut-il pour autant en donner la définition logique du carré ? Est-il capable de dire qu'il s'agit d'un quadrilatère doté de quatre côtés ? Comment définira-t-il ensuite le quadrilatère ? Peut-il retrouver les axiomes d'Euclide par la seule analyse statistique des figures géométriques ? Cette aptitude est celle qui consiste à abstraire la diversité des données de l'expérience de façon à réduire les concepts à leurs unités minimales.

Ce premier type de problème renvoie à une seconde difficulté, qui est celle de la *généralisation*. Les algorithmes bottom-up reposent souvent sur l'usage massif des big data. Ce sont des millions de pages de textes qui sont données, chaque jour, à ChatGPT, pour améliorer ses performances. L'algorithme qui en résulte se comporte comme un être pantagruélique, qui dévore, qui émiette, qui incorpore les savoirs de l'humanité. Or, il est légitime de se demander ce que deviendrait l'intelligence de la machine en présence d'un nombre restreint de données à analyser. L'être humain n'est-il pas capable de raisonner en présence d'un nombre de données finalement très faible, en comparaison des machines utilisées aujourd'hui ?

Cet enjeu est celui qui fait dire, aux technocrates en col blanc, qu'il est grand temps de s'intéresser non pas uniquement aux big data, mais aux quality data. J'aimerais l'illustrer à l'aide d'un exemple. Si je vous montre trois formes que vous n'avez encore jamais vues, puis que je vous demande

de choisir au sein de quatre nouvelles formes celle qui ressemble le plus à celles-ci, il est probable que vous n'ayez pas de difficulté à répondre. Cette propriété est un aspect essentiel de ce qu'on appelle l'intelligence fluide. Contrairement à l'intelligence dite cristallisée, qui reflète la culture générale et le niveau d'éducation, l'intelligence fluide cherche à raisonner sans connaissance préalable significative. Elle fait *tabula rasa* et construit des concepts étape par étape.

Les problèmes de l'abstraction et de la généralisation sont liés intrinsèquement. L'abstraction est en effet une capacité à dépouiller l'expérience de ses disparités pour la ramener à des formes logiques simples. La généralisation, quant à elle, consiste à raisonner à partir d'un petit échantillon de données. Dans un cas comme dans l'autre, l'esprit semble manipuler des concepts intuitifs. Ces concepts ont ceci de particulier qu'ils débordent toujours de la définition logique qu'on leur donne.

Un exemple de concept intuitif très simple est celui d'un « arbre ». Tout être humain bien constitué est en mesure d'identifier un arbre au milieu d'une prairie, mais il serait pourtant bien en peine de donner la définition exacte d'un arbre. L'arbre, en effet, n'est pas « un tronc muni de racines et de branches ». Nombre de plantes respectent ces propriétés sans être des arbres (un plant de tomates, le bambou...). De plus, il existe des arbres munis de plusieurs troncs, tout comme il existe des arbres sans branches. Le concept intuitif « d'arbre » ne tire donc pas son intelligibilité de sa définition linguistique, mais semble relever d'une capacité de jugement pratique.

Pour le dire autrement, l'être humain structure l'information en la ramenant à des objets simples : des chaises, des tables, des carrés, etc. La machine, au contraire, structure l'information en la découpant en un ensemble de points disjoints : les pixels, les instants temporels, etc. Elle tente ensuite après coup de reconstruire des concepts en liant ces points les uns aux autres. Il n'est pas sûr que cette approche *bottom-up* parvienne à se saisir des relations complexes, et intuitives, qui gouvernent la compréhension du tout. Dans les pages qui suivent, je montrerai qu'il existe des tâches d'abstraction et de généralisation, inventées par l'ingénieur François Chollet, sur lesquelles nos meilleurs algorithmes peinent encore à apporter une réponse — et qui sont pourtant naturelles pour l'être humain ⁷. L'objet de ma démarche consistera donc à mettre à l'épreuve la thèse suivante :

Thèse 1. *Ni la programmation top-down ni la programmation bottom-up ne sont capables de former des concepts intuitifs à part entière, mais uniquement de raisonner sur des définitions ou d'opérer des associations statistiques. Il existe un type de jugement qui ne se réduit ni à de purs jugements analytiques ni à de purs jugements synthétiques, mais qui procède par abstraction et généralisation.*

7. Chollet rapporte que près de 85% des problèmes sont résolus, en moyenne, par un être humain (c.f. *ARC-AGI* on <https://arcprize.org/arc>).

Le QI artificiel

Cette difficulté que j'ai épinglée, d'après laquelle les algorithmes peinent à abstraire et généraliser leurs raisonnements, est connue du monde informatique. Elle fait d'ailleurs l'objet de batteries de tests bien spécifiques. On peut penser en particulier au protocole ARC, créé en 2019 par François Chollet, et consistant en un ensemble de problèmes de généralisation semblables à des matrices de Raven (voir figure 4). Dans ce protocole, on présente à la machine deux grilles de pixels : une grille A en input et une grille A' en output. Elle doit comprendre quelle transformation a subie l'image A pour devenir l'image A'. On répète l'opération avec les grilles B et B', puis C et C'. Sur base de ces trois seuls exemples, il lui est demandé de transformer une image D en D'.

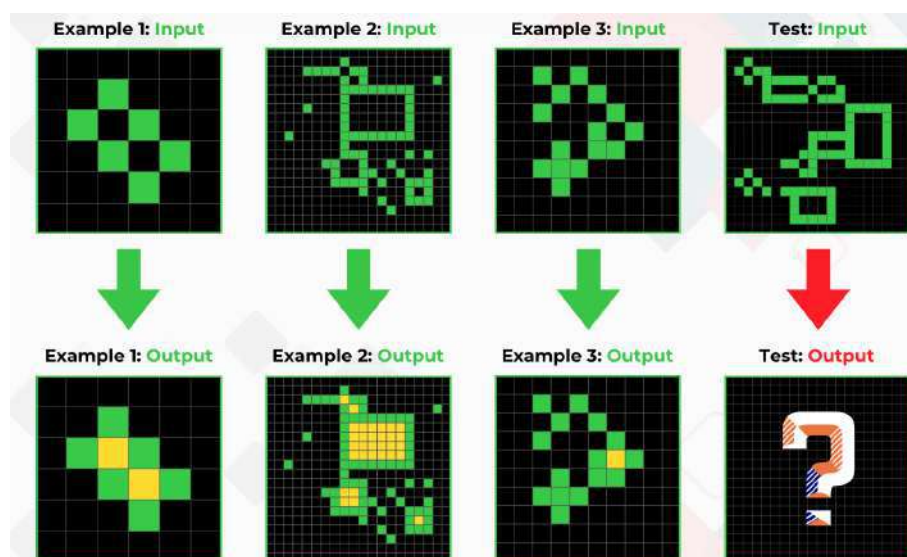


FIGURE 4 – Voici un exemple de test ARC. À partir d'un input donné, la machine doit fournir le bon output. Elle reçoit trois exemples pour comprendre ce qu'elle doit faire (source : <https://lab42.global/arc/>).

La différence entre le problème que je viens d'exposer et ceux qui sont habituellement traités par l'intelligence artificielle, c'est qu'il y a très peu de données disponibles pour résoudre la matrice de Raven. La machine doit performer à partir d'un échantillon de trois exemples seulement. On est bien loin des big data utilisées par des logiciels tels que ChatGPT, qui sont entraînés sur plus de 800 Go de données ! Dans le benchmarking traditionnel, on croule si bien sous le poids exorbitant et quasi prométhéen des données, que toute l'entreprise consiste à les trier, les classer, les éliminer au besoin. Avec le test de QI artificiel, au contraire, on parlera de quality data. Ce qui importe, ce n'est pas le nombre de données, mais la qualité des exemples

présentées et, surtout, la capacité de l’algorithme à généraliser son raisonnement.

Les images à analyser sont le plus souvent des formes géométriques simples, composées d’un petit nombre de pixels, auxquelles on a fait subir des transformations telles que des rotations, des changements de couleurs, des zooms, des additions de pixels, etc. D’après François Chollet, ces exercices mobilisent un certain nombre de fonctions cognitives clefs, telles que celles présentées ci-dessous^[8] :

- **Persistance de l’objet** : Capacité à repérer des objets et à les manipuler sans qu’un changement d’orientation ou de contexte fasse disparaître l’identification de cet objet.
- **Orientation vers d’un but** : Capacité à imputer des intentions ou des buts à certains objets.
- **Arithmétique** : Capacité à trier les objets en fonction de leur forme, de leur couleur, de leur apparence ou mouvement en s’appuyant sur l’addition, la soustraction et la comparaison.
- **Géométrie et topologie** : Capacité à réaliser des manipulations spatiales sur les objets, telles que des rotations, des symétries orthogonales, des translations, des déformations, des répétitions, des associations, etc. Il est possible, en particulier, de déterminer la distance entre deux objets.

On peut constater que ces opérations de pensée sont proches de ce que j’avais en vue lorsque je parlais du jugement intuitif. Il s’agit en effet, pour l’essentiel, de catégoriser l’information sensible de façon à en faire des objets manipulables. Autrement dit, il faut associer une forme géométrique rationnelle à des données disparates. Une étude a d’ailleurs démontré que l’être humain, en moyenne, réussissait 87% des exercices conçus par François Chollet. En comparaison, le meilleur algorithme en septembre 2024 ne dépasse pas la barre des 55%^[9].

Il est donc permis de penser que le problème de généralisation oppose au développement de l’intelligence artificielle des difficultés fondamentales. Il se

8. Ces propriétés sont exposées par l’auteur sur le site internet dédié aux tests ARC (<https://arcprize.org/arc>). Pour de plus amples détails, le lecteur est prié de se référer à sa publication intitulée FRANÇOIS CHOLLET, *On the Measure of Intelligence*, November 5, 2019.

9. Ce nombre était de 30% au début de l’année 2024 et il ne cesse de monter depuis que les informaticiens se sont penchés dessus. Voir à ce titre AYSJA JOHNSON, WAI KEEN VONG, BRENDEN M. LAKE, & TODD M. GURECKIS, *Fast and flexible : Human program induction in abstract reasoning task*

pourrait en effet que les écarts de performance observés entre l'humain et la machine trahissent une lacune dans l'exercice du jugement ; que l'opération qui consiste à appliquer un concept rationnel à une réalité sensible ne soit pas du même ordre que tout ce qu'on a réalisé jusqu'à présent. C'est l'hypothèse que je souhaite envisager dans cette thèse.

Il ne faut toutefois pas exclure que les résultats mitigés obtenus par les méthodes actuelles d'intelligence artificielle aux exercices ARC soient dus à la nouveauté de ces problèmes. Jusqu'à présent, en effet, les programmeurs recourent massivement aux *benchmark* (bases de données standardisées) pour entraîner leurs algorithmes sur des tâches spécifiques, comme l'analyse d'images, les robots conversationnels ou la création de moteurs de recherche. Cette méthodologie s'est avérée si efficace que les concepteurs ne sont guère incités à inventer des algorithmes susceptibles de raisonner dans des situations plus indigentes en termes de données.

Il n'est donc pas interdit de penser que les algorithmes s'amélioreront rapidement sur les tests ARC au cours des prochaines années¹⁰. L'histoire de l'intelligence artificielle a démontré, à maintes occasions, qu'il était périlleux de fanfaronner avec des déclarations du type : « voici ce que la machine ne sait (toujours) pas faire ! » Faut-il rappeler que AlphaGo a été capable de vaincre Lee Soghol en 2016 sans aucune base de données préalable ? Pourquoi les matrices de Raven à la sauce François Chollet ne pourraient-elles pas suivre le même destin ?

Je ne nie pas cela. Mon point consiste plutôt à dire qu'un algorithme vraiment intelligent, qui non seulement remporte l'épreuve de François Chollet dans la pratique — mais aussi dans son esprit — ne devrait pas s'apparenter à un spécialiste froid et sans âme, à un technicien des matrices de Raven, brillant dès qu'on lui donne un problème connu, mais incapable d'appliquer ses formidables méthodes en dehors des ornières étroites sur lesquelles il a été formé.

Ce à quoi l'on aspire fondamentalement, le rêve, le cauchemar et le fantasme de Pygmalion, c'est la création d'une machine capable de trouver des formes abstraites dans des contextes divers et variés. François Chollet précise en outre qu'une intelligence artificielle générale devrait en outre être capable d'exporter ses conclusions et ses méthodes à un nouvel environnement¹¹ :

10. J'aimerais toutefois souligner qu'on a tendance à sous-estimer le temps requis pour surmonter une tâche difficile. En 1958, Herbert Simon affirmait qu'il faudrait dix ans pour qu'un logiciel d'échecs soit capable de vaincre le champion du monde, alors que cela ne s'est pas produit avant 1997. Ce phénomène est résumé par la loi de Hofstadter, qui exprime en substance : « Réaliser une tâche complexe prend toujours plus de temps que prévu, même lorsque vous prenez en compte la loi de Hofstadter. » (voir HOFSTADTER dans *Gödel, Escher, Bach : An Eternal Golden Braid*).

11. Traduit par mes soins depuis l'anglais : « AGI is a system that can efficiently acquire new skills outside of its training data. This means that a system is able to adapt to a new environment that it has not seen before and that its creators (developers) did not anticipate.. » (cf. la page web officiel des tests ARC : <https://arcprize.org/arc>).

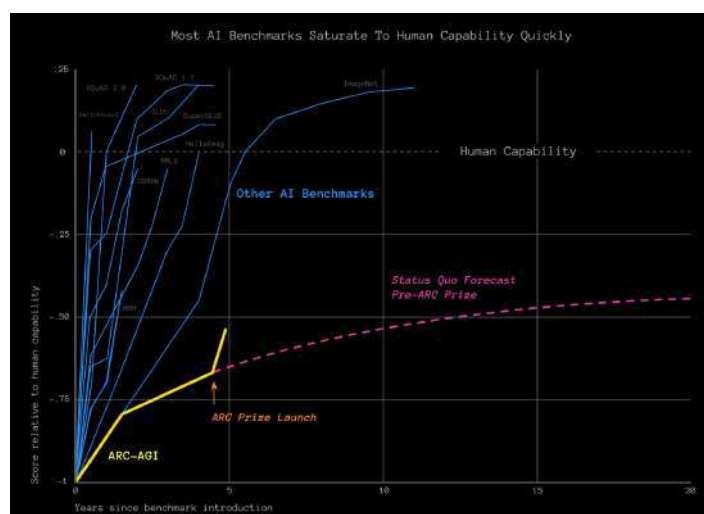


FIGURE 5 – La plupart des algorithmiques finissent par dépasser les capacités humaines sur toutes sortes de tâches. Les progrès face aux tests ARC suivaient une asymptote assez faible, jusqu'à ce qu'un prix de 1,1 million d'euros n'ait été alloué à la tâche (source : <https://lab42.global/arc/>).

L'AGI (Intelligence Artificielle Générale) est un système capable d'acquérir efficacement de nouvelles compétences en dehors de ses données d'entraînement. Cela signifie qu'il est capable de s'adapter à un environnement qu'il n'a jamais vu auparavant et que ses créateurs (développeurs) n'avaient pas anticipé.

Plan de la thèse

La thèse est décomposée en quatre volets. Bien qu'elles suivent un fil conducteur, logique et chronologique, ces parties sont rédigées de telle sorte à être auto-suffisantes. Il vous est donc loisible de vous rendre aux sections qui vous intéressent le plus, de sauter des passages, d'y revenir à votre gré, d'explorer les chapitres comme s'il s'agissait des couloirs d'un astucieux labyrinthe. Les quatre parties en question sont :

1. Le Problème du jugement en philosophie :
De l'analytique au synthétique
2. Révision du concept de jugement :
Vers une théorie Méréologique
3. Le Jugement intuitif face à la Machine :
Extension de la machine de Turing
4. Vers un Monisme à double aspect :

Ce que disent les neurosciences

Je vais résumer ici l'objet de chacune de ces parties. Mais auparavant, j'aimerais attirer votre attention sur un point important : le premier chapitre de chaque partie offre toujours un résumé de ce qui a été dit précédemment, mais aussi de ce qui va être dit. Ces chapitres jouent donc le rôle de points relais. Vous y trouverez du repos, un brin d'humour et quelques prophéties sur la marche future de l'ouvrage. Si votre expérience de lecture s'apparente davantage à un livre policier qu'à la restitution soporifique d'un manuel scolaire ou d'une encyclopédie, j'estimerai avoir au moins eu le mérite d'échapper à cette terrible accusation de Johann Wolfgang von Goethe, lorsqu'il disait à propos des productions intellectuelles de son époque :

Certains livres semblent avoir été écrits, non pour nous instruire, mais pour qu'on sache que l'auteur savait quelque chose.

Partie 1 : Le Problème du jugement en philosophie

Dans la première partie de cette thèse, je vais commencer par analyser la façon dont le concept de *jugement* est défini dans la tradition philosophique. Je suggérerai que l'on trouve, dès l'époque de Platon, une césure entre les jugements analytiques et les jugements synthétiques. La philosophie grecque, en effet, est née de la volonté d'élever la connaissance, empêtrée dans des observations sensibles, vers plus de rationalité. Elle pose d'un côté l'existence d'une vérité fondée sur l'expérience et, de l'autre côté, la présence d'une vérité fondée sur les idées. La dialectique socratique, à ce titre, peut être vue comme une tentative d'élucider, par la seule force du langage, l'essence du courage, de la science, de l'amour, de la sagesse, de l'art et des autres concepts que nous utilisons au quotidien.

« Si l'on nous interrogeait sur une chose banale et commune, par exemple sur ce que peut être la boue », explique Socrate dans un dialogue l'opposant à Théétète, « et que nous répondions qu'il y a la boue des potiers, la boue des constructeurs de fours, la boue des briquetiers, ne serions-nous pas ridicules ? » Par cette affirmation, le philosophe d'Athènes exprime clairement l'idée que la boue n'est pas la somme statistique, la synthèse grossière, de toutes les formes de boue possibles et imaginables. La boue est « de la terre délayée avec de l'eau ». C'est une idée dotée d'une définition rationnelle, d'une essence immuable. Il n'est pas question de la confondre avec les apparences diverses, et forcément contingentes, qui échoient à l'expérience sensorielle.

Ce qui est intéressant, c'est que Platon ne se contente pas d'une division strictement binaire, en opposant le monde sensible (synthèse) au monde intelligible¹² (analyse). Il envisage également la présence d'une forme de

12. Notons que l'expression « monde intelligible » n'existe pas comme telle chez Platon,

jugement située à l'exact milieu entre le ciel des définitions rationnelles et la terre ferme de l'expérience. Il donne à celle-ci le nom de « jugement vrai sans justification ». Ce qu'il entend par là, c'est une certaine aptitude à créer le concept de ce qui est « dur », « mou », « froid », « chaud » ou « boueux », en rassemblant sous une même unité linguistique toute la diversité des choses dures, molles, froides — mais sans en donner pour autant une définition rationnelle. Je tenterai de montrer dans cette thèse que ce jugement sans justification est la forme embryonnaire de jugement intuitif qui fait l'objet de mon enquête philosophique.

Bien que Platon dédie plusieurs passages à ce « jugement sans justification », il ne l'impute pas à une faculté humaine particulière. Précisons ce point. Contrairement au jugement rationnellement justifié, qui est le propre de la raison, et contrairement à l'expérience, qui est le propre de la sensation, le jugement intuitif n'est rien de plus qu'une aptitude générale de l'âme humaine. C'est ce que Platon exprime par la bouche de ses personnages : « Je ne sais que dire, sinon qu'à mon avis, il n'y a pas du tout d'organe spécial pour ces [jugements], comme il y en a pour les autres : c'est l'âme elle-même et par elle-même qui, selon moi, examine les notions communes en toutes choses¹³. »

Tout se passe donc comme si Platon reconnaissait l'existence d'une forme de jugement située entre l'expérience et la raison, mais qu'il ne lui donnait pas toute l'importance qu'elle mériterait. Cette omission spectaculaire vient peut-être de son idéal de rationalité ? La marotte de Socrate, dans tous les dialogues où on le voit apparaître, est toujours la même. Qu'il s'agisse de parler de la vertu, de l'art de commander ou de la façon de siphonner plusieurs amphores de vin, ses discours tentent sans cesse de rapporter tous les concepts intuitifs que nous nous formons à propos du monde à des concepts rationnellement argumentés. Non seulement la boue est « de la terre délayée avec de l'eau », mais la science est « le jugement vrai muni de raison ».

Il y a un mouvement, dans les balbutiements de la rationalité occidentale, qui semble avoir mis entre parenthèses la légitimité du raisonnement intuitif, défini comme la faculté de juger du vrai sans s'appuyer ni sur la logique formelle ni sur de simples lois statistiques. Ce mouvement, prolongé par la science moderne, est ce que j'appellerai dans cette thèse « L'Oubli de l'Entendement ». Le mot « Entendement » figurant dans cette expression désigne, en vieux français, le fait de saisir quelque chose par l'intelligence, de pouvoir en juger, sans pour autant détenir une explication purement rationnelle. Il s'agit d'un intermédiaire entre la simple perception sensorielle (ouïr quelque chose) et le raisonnement logico-déductif (comprendre quelque chose). Il est d'ailleurs intéressant de remarquer que le mot « entendement »

mais est d'invention plus récente. On trouve en revanche l'expression « lieu intelligible » (*topos noetos*)

13. *Théétète*, 186b-186c.

est tombé en désuétude dans la langue française. On parle aujourd'hui d'entendre un bruit ou un son, et non d'entendre les mathématiques ou la poésie. Cette omission linguistique de « l'entendement » accompagne l'omission proprement philosophique du jugement. Elle participe de la représentation scientifique du monde moderne, d'après laquelle toute vérité peut se réduire à des calculs rationnels.

À l'issue de cette première partie sur Platon, je montrerai en quoi le *jugement synthétique a priori* d'Emmanuel Kant hérite de cette tension historique opposant le jugement analytique (définition rationnelle) au jugement synthétique (issu de l'expérience). J'utiliserai en outre la théorie kantienne comme point de départ pour questionner les limitations de l'intelligence artificielle. Il s'agira d'identifier une première ligne de fracture essentielle entre la cognition humaine et le fonctionnement de nos algorithmes actuels : la façon dont l'information, ou les *data*, sont structurées.

Je suggérerai que les algorithmes actuels découpent l'information en unités irréductibles (les pixels, les bits...), puis cherchent à associer ces unités minimales d'après des lois statistiques de façon à identifier des structures (reconnaître l'image d'un arbre, comprendre le sens d'une phrase, etc.). L'être humain, au contraire, analyse l'information dans sa globalité ; plutôt que de la découper en pixels indivisibles, il la voit au travers de structures préalables. Ces structures sont *a priori* dans le sens précis où, en l'absence de ces structures, il n'y aurait pas de perception possible. Autrement dit c'est parce que nous avons le concept de carré, de triangle, d'espace, de table, que nous pouvons nous représenter une table carrée au milieu de la pièce.

J'illustrerai cette différence existant entre la cognition humaine et les algorithmes au moyen de plusieurs illusions perceptives, notamment l'illusion de Kanizsa et celle de Ponzo. Dans l'illusion de Kanizsa (figure 3.2), des sphères sectionnées sont représentées, et agencées de façon à suggérer l'existence d'un triangle. Aucun segment de ce triangle n'est dessiné explicitement, mais un esprit humain ne peut se débarrasser de l'impression, quelque peu gênante, que le triangle existe bel et bien. Cette illusion illustre l'idée que l'information n'est pas traitée de façon purement objective, comme le ferait un algorithme, en associant à chaque point de l'espace un pixel de couleur ; au contraire, l'image dans son ensemble nous apparaît conformément à certaines structures préalables. C'est parce que nous disposons — préalablement à la contemplation de l'image — du concept de triangle, que nous voyons un triangle. On peut donc penser que l'image perçue par l'esprit humain suit un découpage qui dépend de nos concepts *a priori*.

Dans l'illusion de Ponzo (figure 3.4), quelque chose d'encore un peu différent se produit. On y voit représentés deux lignes horizontales. Bien qu'elles aient strictement la même taille, celle du haut nous semble plus longue. Cette illusion d'optique est créée par la présence, dans l'arrière-plan, de deux autres lignes — qui suggèrent l'existence d'une perspective. Si vous retirez ces lignes en arrière-plan, l'illusion s'évanouit. Contrairement à un

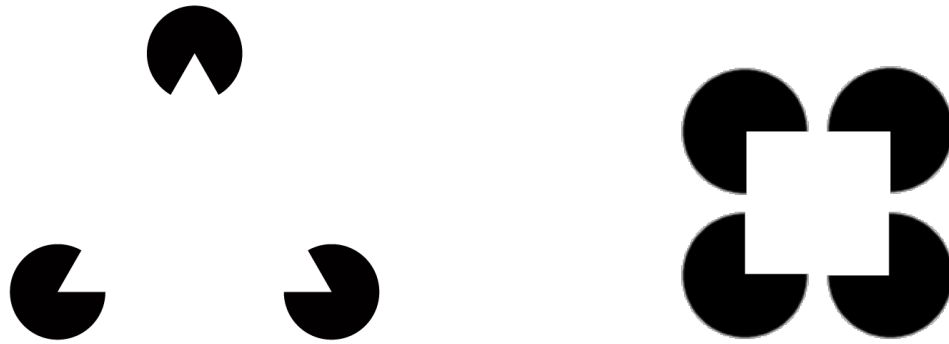


FIGURE 6 – Dans les illusions de Kanizsa, une figure géométrique apparaît dans le blanc de la page. D’un point de vue kantien, cela illustre que c’est le monde sensible qui s’adapte à nos catégories, et non l’inverse.

algorithme, notre cerveau ne peut pas ne pas être victime de l’illusion. Cela suggère que notre perception ne se contente pas de regarder les objets individuellement, mais qu’elle est influencée par l’agencement des objets les uns par rapport aux autres. Il y a un phénomène global, dans la perception, qui ne semble pas répondre aux logiques de programmation présidant aujourd’hui à l’analyse des images. Ce phénomène de perception global, du reste, est influencé par nos concepts préalables d’espace et de tridimensionnalité. Fait notable : l’illusion est moins forte chez les enfants qu’auprès des adultes ; elle est également moins forte chez les sujets souffrant de schizophrénie¹⁴.

Partie 2 : Révision du concept de jugement

Dans cette deuxième partie de la thèse, il s’agira de redéfinir le concept de jugement. Bien que la théorie de Kant constitue, comme je l’ai suggéré, un bon point de départ pour critiquer l’intelligence artificielle, elle ne peut pas être conservée en l’état. L’histoire récente des mathématiques et de la physique semble avoir en effet porté un coup rude à l’idée kantienne d’après laquelle nous disposons de structures *a priori*, qui ne dépendraient pas de l’expérience. À titre d’exemple, on peut penser à la géométrie. Kant pensait que les objets géométriques, et plus généralement notre représentation de l’espace en trois dimensions, était une forme *a priori* de la cognition hu-

14. Pour l’effet de l’âge, se référer à *Aging reduces center-surround antagonism in visual motion processing*. Betts LR, Taylor CP, Sekuler AB, Bennett PJ. *Neuron*. 2005 Feb 3;45(3):361-6. doi : 10.1016/j.neuron.2004.12.041. Pour l’impact de la schizophrénie, voir notamment YANG, EUNICE ; TADIN, DUJE ; GLASSER, DAVIS M. ; HONG, SANG WOOK ; BLAKE, RANDOLPH ; PARK, SOHEE (2012-11-15), *Visual Context Processing in Schizophrenia*, *Clinical Psychological Science*. 1 (1) : 5–15. Notons que ces études que j’évoque ici ne concernent pas particulièrement l’expérience de Ponzo, mais l’expérience de Chubb, consistant à distinguer un objet placé dans un arrière-plan confus.

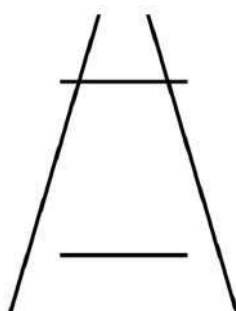


FIGURE 7 – Illusion de Ponzo : Les deux lignes horizontales ont la même taille, mais celle en haut de l'image semble plus grande.

maine : l'intuition de l'espace. D'après lui, les vérités mathématiques étaient fondées sur cette intuition ; cela veut dire qu'en l'absence de cette intuition, il ne serait pas possible de construire l'échafaudage théorique de la géométrie.

Or depuis la publication des œuvres de Kant, bien de l'eau a coulé sous les ponts. Des mathématiciens tels que Lobatchevski ou Riemann ont montré qu'il était possible de construire des vérités mathématiques à propos de géométries non euclidiennes. Dans ces géométries, la somme des angles d'un triangle ne fait pas 180 degrés. Les notions de droites, de parallèles ou de « chemin le plus court » ne sont plus tout à fait intuitives. Tout se passe donc comme si nos intuitions *a priori* n'existaient pas de façon totalement indépendante de l'expérience du monde. C'est parce que nous faisons l'expérience d'un monde en trois dimensions que nous acquérons une intuition tridimensionnelle de l'espace. Mais si nous évoluions dans un milieu aquatique, où les objets se déformaient d'après les courants, notre intuition de l'espace ne serait-elle pas différente ? Le « chemin le plus court » entre deux points serait-il toujours une ligne droite, ou épouserait-il la direction des marées ?

En tenant compte de ces importantes critiques, j'entreprendrai de reconstruire le concept de jugement de façon à garder la tension entre les jugements analytiques et les jugements synthétiques, mais sans présupposer l'existence de structures innées. Je plaiderai en faveur d'une interprétation circulaire entre l'intuition et l'expérience, consistant à dire que (1) les structures sont ce sans quoi l'information perceptive ne serait pas possible ; (2) mais que ces structures sont, d'autres part, influencées par notre expérience perceptive. Il s'agit d'un juste milieu entre une théorie kantienne de la cognition humaine et une théorie empiriste.

Je suggérerai en outre qu'il existe trois niveaux de traitement dans l'intelligence humaine :

1. **La rationalité** : Ce niveau comprend les *concepts rationnels*, définis

à l'aide de critères rigoureux et vérifiables. À titre d'exemple, l'idée du « carré » est un concept rationnel ; c'est très exactement « une figure plane dotée de quatre angles droits ». La « justice », une « maison » ou un « être humain », en revanche, ne sont pas des concepts rationnels car leurs définitions souffrent d'un certain flottement. Il serait impossible de définir précisément ce qu'est un être humain à l'aide de mots. On peut certes dire qu'il s'agit « d'un quadrupède doué de telles et telles propriétés (le langage, la raison, un certain code génétique...) » mais si la définition est trop restrictive, certains humains risquent de ne pas la satisfaire ; tandis que si la définition est trop lâche, des mammifères non-humains pourraient être inclus à tort. Les concepts rationnels se limitent donc à la sphère idéale des mathématiques et de la logique.

2. **La Sensation** : Ce niveau comprend les *percepts sensibles*, définis comme un certain regroupement d'objets présentant, entre eux, des similarités. Ces percepts ne sont pas des concepts à part entière, dotés d'une définition. Il s'agit seulement de schémas d'analogies, réalisés le plus souvent de façon automatique et inconsciente. Un bon exemple est la contemplation d'une peinture abstraite. Lorsque vous la contemplez, vous opérez un découpage consistant à distinguer certaines formes de couleurs les unes par rapport aux autres. Ces formes ne disposent pas de définitions strictes. Vous ne pouvez pas dire que c'est « une maison », un « être humain » ou même « une couleur rouge ». Ces formes n'existent qu'en tant que contrastes ; elles sont donc le fruit de l'expérience sensible, dépouillée de toute structure rationnelle rigide.
3. **La Perception (ou l'Entendement)** : Ce niveau comprend les *concepts intuitifs*, tels que la « justice », une « maison » ou un « être humain ». Il s'agit bien de concepts puisqu'on peut les définir au-delà de notre expérience immédiate. Lorsque je contemple une « maison », je sais que l'exemplaire que j'ai sous les yeux n'est pas la seule maison possible et imaginable. Il ne s'agit que d'un cas singulier. Bien que je sache cela, il me serait difficile d'expliquer de façon précise ce qu'est une maison dans le sens le plus général du terme. Si on me pose la question « qu'est-ce qu'une maison ? », mon premier réflexe consistera à énumérer différentes maisons : « Les bungalows, les villas, les cabanes, les appartements, tout cela sont des maisons ». Ce faisant, je ne réponds pas vraiment à la question. Mais est-il possible de donner une définition exacte de la maison ? Sans doute pas. C'est pourquoi le concept de maison est dit intuitif. On ne peut se le représenter qu'au travers d'un archétype (une maison type), mais on sait que la maison ne se réduit pourtant pas à cet archétype. Dans ma thèse, je défendrai que les concepts intuitifs sont le liant entre le niveau sensible de l'expérience et le niveau abstrait de la raison.

La théorie tricamérale de la cognition dont je me fais l'avocat peut être

vue comme une extension des systèmes 1 et 2 de David Kahneman, qui a reçu le prix Nobel d'économie en 2002. David Kahneman insiste sur l'existence, dans le chef de l'esprit humain, de deux systèmes de traitement de l'information : (a) un système automatique, rapide, statistique, qui ressemble au fonctionnement des réseaux de neurones et (b) un système conscient, lent, rationnel, qui s'apparente davantage à la programmation top-down. Ces deux systèmes de Kahneman ne mobilisent pas les mêmes régions neuronales et ne répondent pas des mêmes principes, mais ils sont tous deux importants pour la prise de décision humaine. Dans son livre intitulé *Système 1 et Système 2 : les deux vitesses de la pensée*, Kahneman donne la description suivante¹⁵ :

Le Système 1 opère de manière automatique et rapide, avec peu d'effort et sans s'accompagner du sentiment d'exercer un contrôle volontaire. Le Système 2 alloue les ressources attentionnelles aux activités mentales exigeantes, y compris les calculs complexes. Les opérations du Système 2 sont souvent associées à l'expérience subjective d'exercer son libre arbitre et de faire un effort de concentration.

Je suggérerai qu'il existe un troisième système, situé au milieu des deux premiers. Ce Système 3, que la philosophie appelle traditionnellement du nom de *l'Entendement*, est une opération certes consciente, mais pas totalement justifiable. Elle consiste à penser les concepts d'une façon qui ne soit ni statistique ni rationnelle, mais intuitive. Cette intuition se fait via l'élaboration d'une image mentale, un schème. Cette image est certes inspirée par des représentations concrètes, mais elle désigne quelque chose qui dépasse toutes les représentations possibles. Qu'est-ce que je veux dire par là ? Lorsque vous vous représentez l'idée d'un triangle ou d'un arbre, vous voyez peut-être un triangle isocèle ou la forme typique d'un chêne, mais cette image mentale que vous produisez dans votre esprit ne désigne pas un triangle isocèle ou un chêne en particulier ; elle désigne l'idée abstraite du triangle ou de l'arbre. C'est un archétype universel, une représentation qui vise autre chose que sa représentation.

Je ferai l'hypothèse que le niveau supérieur de la cognition (la raison) et le niveau inférieur de la cognition (la sensation), peuvent être imités par les algorithmes actuels, mais non le niveau intermédiaire (la Perception ou l'Entendement). Le niveau supérieur s'apparente en effet à la programmation top-down dont j'ai parlé dans l'introduction. Il consiste à représenter l'information sous forme de symboles (des nombres, des caractères algébriques) et à les manipuler d'après certaines règles logiques de façon à arriver à une conclusion rationnellement valide. C'est ce que font les robots d'Isaac Asimov lorsqu'ils jugent, d'après les lois de la robotique, si un comportement doit être qualifié de moral.

15. Cette citation est traduite de l'anglais par mes soins.

Le niveau inférieur de la cognition, quant à lui, s'apparente à la programmation bottom-up. Il consiste à partir des données diverses de l'expérience et à les associer de façon à construire des concepts. Les algorithmes utilisés à cette fin reposent massivement sur les statistiques : les réseaux de neurones ne sont d'ailleurs rien d'autre qu'une régression statistique multivariée. Il en découle des réponses adaptatives, flexibles, sensibles au contexte, mais susceptibles de commettre des biais. Un exemple peut être trouvé dans ChatGPT, qui tend à produire les réponses les plus statistiquement valides, même lorsque celles-ci sont inadéquates. Il y a d'ailleurs une armée de mains invisibles — des utilisateurs pour la plupart, mais pas que — qui donnent du feedback à l'algorithme pour en corriger itérativement les écarts.

Le niveau intermédiaire (la perception), pour finir, procède de façon holistique. C'est-à-dire qu'il applique les structures conceptuelles de la raison aux données de l'expérience. Cette démarche n'est ni top-down, ni bottom-up, mais circulaire. Cela veut dire que les données ne sont rendues visibles qu'aux travers des structures et que les structures, à leur tour, n'existent qu'au contact des données. Il s'ensuit que la perception n'est pas comparable, ni à l'association statistiquement aveugle des réseaux de neurones, ni à la séquentialisation étroite de la raison ; elle naît au contraire des contrastes. C'est en portant son attention sur les différences et les similitudes que l'esprit humain découpe le tout perceptif en unités individuelles. Les objets de perceptions sont le fruit d'un principe de différenciation, qui utilise les concepts pour séparer les contenus sensibles selon certains motifs.

Bien que ce niveau intermédiaire puisse sembler comparable au fonctionnement des réseaux de neurones, je tenterai de montrer, exemples à l'appui, qu'il repose sur deux intuitions : l'intuition du tout et l'intuition de l'infini. Ces intuitions que j'introduis remplacent les intuitions de l'espace et du temps que l'on trouvait dans la théorie kantienne. Je montrerai que le Tout et l'Infini interviennent dans notre représentation spatio-temporelle des phénomènes, qu'ils supportent en quelque sorte ces derniers, sans supposer néanmoins que l'espace et que le temps se présente sous une forme euclidienne. On pourrait dire, en bref, qu'il s'agit des structures les plus fondamentales de notre perception humaine du temps et de l'espace. Je les caractérise ici brièvement :

- **L'intuition du tout** : Cette intuition est ce qui nous permet de saisir l'information de façon globale, et non séquentielle. Lorsque cette intuition est appliquée à notre expérience du temps, elle nous permet de comprendre que l'instant présent n'existe pas sans référence à un instant passé et à un instant futur. Ce n'est qu'une fois que l'instant en question s'est écoulé que nous pouvons, rétrospectivement, le prendre pour objet de pensée, si bien que le processus de pensée se déploie dans un va et vient continu entre « ce que je suis en train de penser maintenant » et « ce que je pense que j'étais en train de

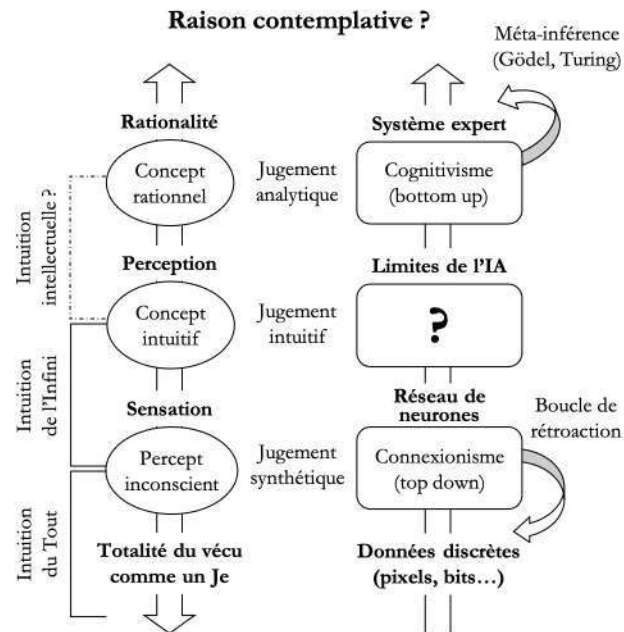


FIGURE 8

penser il y a un instant ». Autrement dit, notre perception du temps est globale.

- L'intuition de l'infini :** Cette intuition est ce qui nous permet de comprendre et de donner un sens à l'ambiguïté, en créant un tissu d'association entre nos concepts intuitifs. Elle est à l'œuvre chaque fois que l'on rapporte un objet à un concept. Lorsque l'on dit « cet arbre est un bouleau », on associe et on différencie partiellement cet arbre que nous avons sous les yeux de l'infinité des bouleau possibles et imaginables. L'arbre singulier ne tire son intelligibilité que de l'infinité des concepts possibles auxquels il se rapporte. Inversement, le concept de bouleau ne tire son intelligibilité que de l'infinité des arbres singuliers possibles auxquels il se rapporte. C'est pourquoi on en apprend parfois autant en disant « cet arbre est un bouleau » qu'en disant « les bouleaux sont comme cet arbre ». Cette commutativité entre le sujet et le prédicat, spécifique au langage humain, nous permet de comprendre le lien entre les concepts universaux (bouleau) et les objets singuliers (cet arbre).

La figure 8 récapitule cette théorie tricamérale de la cognition dont je me fais l'avocat. On observe que l'esprit humain, représenté à gauche de ce schéma, est doté de trois facultés : la Sensation, la perception (ou Entende-

ment) et la Raison. Celles-ci partent du « Je pense » cartésien, cet état de perception originel où l'expérience d'être quelque chose plutôt que rien est donnée comme une totalité. Dans cet état, il n'y a pas de différence entre le sujet et l'objet, entre la pensée, la sensation, le vécu et le monde extérieur. Seul est ce qui est senti, pensé, vécu à l'instant présent. L'esprit humain va toutefois décomposer cette unité de « l'instant présent » pour y faire apparaître des objets indépendants. Ces objets évoluent au cours du temps, ils changent, ils se transforment.

La décomposition de la totalité du vécu en objets individuels n'est pas réalisée d'un seul coup ; c'est un processus qui s'opère par des va-et-vient successifs entre les trois facultés que sont la Sensation (percepts inconscients), la Perception (concepts intuitifs) et la Rationalité (concepts rationnels). Ces niveaux s'appuient l'un sur l'autre : les objets existent au travers du tout et le tout existe au travers des objets. Le fait de passer d'un niveau inférieur de cognition à un niveau supérieur, en partant par exemple du percept pour en faire un concept, est précisément ce que j'ai appelé le *jugement*.

Il existe donc trois types de jugements. Ceux qui consistent à passer de la totalité du vécu (le Je pense) aux percepts inconscients (la Sensation) sont les jugements synthétiques et reposent sur l'intuition du Tout. Les jugements qui permettent de rassembler les percepts (la Sensation) pour en faire des concepts intuitifs (la Perception) sont les jugements intuitifs et reposent sur l'intuition de l'Infini. Enfin, les jugements permettant d'abstraire les propriétés essentielles des concepts intuitifs (la Perception) pour les manipuler de façon logico-déductive (la Raison) sont dits analytiques et reposent sur une hypothétique intuition intellectuelle (attestée chez Platon mais contestée par Kant, pour qui la Raison spéculative a perdu tout contact avec le monde sensible de l'expérience).

La machine, représentée à droite du schéma, fonctionne différemment de l'esprit humain. Elle ne part pas d'un *Je pense* originel, mais d'un ensemble préalable d'unités discrètes indépendantes (pixels, points, nombres). Les algorithmes ont pour tâche d'associer ces unités discrètes pour reconstituer une vision globale dans un mouvement purement ascendant (qui part des parties et s'élève vers le tout). La machine ne possède donc que deux niveaux de traitement : (1) celui des réseaux de neurones, qui correspond chez l'être humain au niveau de la Sensation (paradigme de programmation bottom-up) et (2) le niveau des systèmes experts correspondant à la Rationalité (paradigme de programmation top-down).

Partie 3 : Le Jugement intuitif face à la Machine

Dans cette troisième partie, il s'agira de se demander si l'intuition de l'infini et l'intuition du tout, dont j'ai dit qu'ils étaient les ingrédients essentiels du jugement humain, peuvent être implémentées à l'aide d'une machine de Turing différente de celles que nous utilisons aujourd'hui ? La machine

de Turing en effet constitue le modèle par excellence de l'ordinateur. Elle consiste en un ruban de taille infinie, sur lequel une tête de lecture inscrit des suites de symboles du type 1 ou 0. Cette tête, qui n'est rien d'autre que le stylet sacré de l'informatique, peut lire les symboles existants, les effacer, en écrire de nouveaux et se déplacer autant qu'elle le souhaite le long du ruban.

La question que j'aborderai dans ma thèse consiste à se demander si une machine de Turing serait capable de construire des catégories intuitives, telles que les catégories d'une table, d'un livre, d'un bouleau ? Dans le cas d'une réponse affirmative, quelles seraient les différences entre ces catégories et celles manipulées par l'esprit humain ?

Je suggérerai que la machine de Turing se heurte à deux obstacles théoriques.

- **Commutativité sujet-prédicat** : Une machine de Turing est définie par des assignations univoques. Lorsqu'elle stipule $A = a$, en écrivant un certain symbole donné sur la bande de lecture, cela signifie que la valeur a est assignée à la variable A . Or dans les catégories sensibles de l'esprit humain, on a vu que la relation de prédication est commutative. La proposition « ceci est un arbre » nous en apprend tout autant sur la nature du « ceci » qui se tient face à nous sur ce que sont les arbres en général. La relation d'égalité $A = a$ n'est donc pas stricte dans la connaissance sensible ; elle se contente d'opérer un appariement, une comparaison, entre les deux termes. Les catégories que nous manipulons ne sont pas des classes prédéfinies, qui existeraient quelque part dans le ciel des idées, et dont il suffirait de nous saisir. Il s'agit au contraire d'un processus mouvant, évoluant au gré de notre contact avec l'expérience sensible¹⁶.

16. Ce problème de commutativité sujet-prédicat explique pourquoi la plupart de nos concepts ne peuvent pas recevoir une définition exacte. Il est d'ailleurs important de faire remarquer que la science, malgré sa rigueur, n'échappe pas au phénomène. Les notions mobilisées en mathématiques et en physique ne sont pas définies de façon univoque, mais s'éclairent les unes les autres. Un exemple d'importance pour notre propos est le concept de « calculabilité » en informatique théorique. Bien que l'on saisisse intuitivement que la calculabilité soit la capacité à déterminer, au moyen d'un algorithme ou d'une procédure de raisonnement systématique, la réponse à une question donnée, il est légitime de se demander ce qu'est un « algorithme » ou une « procédure de raisonnement systématique » ? La thèse de Church-Turing stipule que ces algorithmes et procédures de raisonnement peuvent être exprimés à l'aide de plusieurs formalismes de puissance équivalente : la machine de Turing, le lambda-calcul d'Alonzo Church ou le modèle des fonctions générales récursives de Herbrand-Gödel (voir à cet effet l'annexe 3). Chacun de ces modèles donne donc un éclairage au concept de calculabilité ; mais non seulement cet éclairage n'épuise pas totalement la notion de calculabilité, mais il est en outre nécessaire d'avoir une notion préalable de calculabilité pour comprendre la pertinence de ces modèles. La compréhension humaine de ce qui est calculable, loin de répondre à une définition stricte du type $A = a$, se déploie dans un réseau de concepts interconnectés.

- **L’Idée de l’infini** : L’être humain, lorsqu’il construit un percept aussi banal que celui d’un « verre » ou d’un « arbre », manipule l’idée de l’infini. Il est en effet capable de synthétiser sous une même image mentale un ensemble infini dénombrable de verres ou d’arbres singuliers. De façon analogue, il peut énumérer les nombres $1, 2, 3, \dots$ jusqu’à se faire asymptotiquement l’idée d’un ensemble de nombres $\mathbb{N}$ contenant tous les autres. Cette idée de l’infini est non représentationnelle, c’est-à-dire qu’on ne peut pas se la représenter explicitement de façon sensible, mais elle n’en est pas moins manipulable par notre esprit. Autrement dit, on peut penser cet ensemble de nombres $\mathbb{N}$, mais non pas en énumérer le contenu. Le problème, c’est qu’il n’est pas sûr qu’une Machine de Turing puisse manipuler des idées non représentationnelles, et moins encore l’idée de l’infini. Elle possède certes un ruban de taille infinie sur lequel inscrire les nombres $1, 2, 3, \dots$, mais peut-elle se faire une idée de ce qui existe au bout du ruban ? Peut-elle se représenter que le symbole fini $\mathbb{N}$ désigne un objet infini de par ses contours ?

Dans le cadre de ma thèse, je ferai l’hypothèse que chacun de ces deux problèmes, celui de la commutativité sujet-prédicat et celui se rapportant à l’idée de l’infini, correspondent respectivement à l’intuition du tout et à l’intuition de l’infini. Ces intuitions sont l’enclume sur laquelle l’esprit humain, muni du marteau de l’expérience, peut forger les catégories. Sans cette enclume, l’expérience brasserait de l’air. Il lui faut une résistance, une structure solide, pour donner naissance à des objets. L’intuition, et ses facette, est donc la faculté à se saisir des objets, non pas à l’aide d’un raisonnement logico-déductif traçable, qui procède linéairement des prémisses vers les conclusions, mais de façon globale. Si l’on en croit l’étymologie latine, *intuition* viendrait d’ailleurs du mot *intuitio*, qui signifie « l’acte de voir d’un seul coup d’œil. » Il y a une tendance, en philosophie, à opposer le concept d’intuition à celui de raison. Il suffit de penser par exemple à cette phrase de Schopenhauer : « Les concepts forment une classe spéciale de représentations, entièrement distinctes des représentations intuitives dont il a été question jusqu’ici, car elles n’existent que dans l’esprit humain¹⁷ »

Je défendrai que les intuitions du tout et de l’infini ne peuvent pas être modélisées à l’aide d’une machine de Turing classique, mais requièrent une machine de Turing transfinie. Ce qui différencie la machine transfinie de la machine de Turing classique ne repose donc que sur un seul ingrédient : compter l’infini. Concrètement, cela veut dire que la machine transfinie possède un ruban de lecture infini et qu’elle peut atteindre la « fin de ce ruban ». Je montrerai que cette extension des machines de Turing est analogue, quant à son concept, à une machine capable de penser le continu, plutôt que de

17. Schopenhauer A., *Le monde comme volonté et représentation*, Chapitre 9

fonctionner de façon discrète. Pour une machine continue, une droite est infiniment divisible, tandis que pour une machine classique, il existe des unités de mesure atomiques en-deçà desquelles on ne peut plus rien calculer. Il est vraisemblable que l'esprit humain puisse penser le continu, si bien que la machine de Turing transfinie, en ce qu'elle modélise des opérations continues, capture sans doute un aspect réel de notre cognition humaine.

Partie 4 : Vers une théorie générale de l'esprit humain

La dernière partie de cette thèse partira du constat suivant : si l'esprit humain possède l'intuition de l'infini et l'intuition du tout, et si ces intuitions ne peuvent être modélisées qu'au moyen d'une machine de Turing transfinie, comment se fait-il que notre cerveau soit, quant à lui, de taille finie ? Il faut admettre, pour commencer, que la machine de Turing transfinie soit un modèle de l'intuition, mais non pas un modèle du cerveau — dont la mécanique est strictement finie. Il faut en outre se demander si la cognition vient bel et bien du cerveau ? Ou si l'on ne fait pas erreur en pensant que tout ce qui relève de l'esprit doit pouvoir s'expliquer, en dernier ressort, par un phénomène physiologique apparaissant au sein de notre boîte crânienne ?

Face à cette épineuse question du rapport entre le corps et l'esprit, je défendrai un mystérianisme perspectiviste. Le mystérianisme, en philosophie, souscrit à l'idée qu'il existe des mystères, dans la nature, que l'esprit humain n'est pas en mesure d'élucider. Le philosophe McGinn, par exemple, soutient que la conscience fait partie de ce mystère. J'aimerais marcher sur ses brisées, mais à quelques distinctions près. Je ne dis pas que la conscience ou que l'esprit, en soi, soient inexplicables. Je dis que le rapport entre l'esprit et le corps échappe à toute modélisation. Pour être plus exact, le mystérianisme dont j'entends me faire l'avocat tient en trois points :

1. Le fonctionnement du cerveau peut être modélisé dans une approche neurologique.
2. Le fonctionnement de l'esprit humain, entendu dans une perspective introspective, peut être modélisé dans une approche psychologique¹⁸.
3. Il n'existe pas de théorie permettant de comprendre le rapport entre le modèle biologique du cerveau et le modèle psychologique de l'esprit.

18. Ce modèle introspectif de l'esprit humain consisterait à expliquer la dynamique interne du psychisme depuis un point de vue subjectif et phénoménologique. On en trouve des exemples dans la philosophie de Husserl, qui étudie les principes d'après lesquels des objets se constituent dans la conscience. On peut également penser à la philosophie kantienne, qui n'est pas fondamentalement éloigné d'une théorie structurelle de l'esprit humain, ou même à la topic freudienne, qui tente de donner sens à des concepts tels que le Moi.

J'entends déjà s'élever une importante objection. Pourquoi le lien entre le corps et l'esprit, questionnera-t-on, résiste-t-il à toute théorisation ? Il n'est pas si facile de dire, d'une question scientifique ou philosophique, qu'elle n'admet pas de solution. Nul en effet ne peut prédire la marche future de la connaissance, ni la façon dont va évoluer notre représentation même du problème du corps et de l'esprit. Pour vous donner un exemple historique, il semblait inconcevable aux anciens que l'Univers puisse avoir une autre forme que celle d'une sphère ou que la Terre puisse être ailleurs qu'au centre du monde connu. En dépit de ces préconceptions, nous voici aujourd'hui avec des théories où la Terre occupe le bord de la galaxie, au sein d'un Univers en expansion infinie, et dont la forme suit des lois défiant la géométrie euclidienne. Or cette science dont je parle, en dépit de ses progrès, et aussi sublime qu'elle nous paraisse aujourd'hui, en est encore à ses balbutiements. Des choses aussi incompréhensibles que la conscience, aussitôt débarrassée de ses préjugés, ne pourraient-elle pas, dès lors, recevoir une explication dans les siècles à venir ?

Le mystérianisme que je vais défendre consistera à penser que si la relation entre le corps et l'esprit est énigmatique, ce n'est pas en raison d'une lacune au niveau de nos connaissances, mais c'est parce que les mots « corps » et « esprit » désignent une seule et même réalité. Cette réalité, hélas, nous ne pouvons pas la voir dans son unité. Elle se présente à nous comme un cylindre se présenterait à un savant vivant dans un monde en deux dimensions. Pour cet être, le cylindre apparaîtrait tantôt sous la forme d'un cercle, tantôt sous la forme d'un rectangle, selon qu'il soit projeté dans le monde en deux dimensions d'après une projection longitudinale ou transversale. Autrement dit, notre être bidimensionnel verrait deux objets différents, un cercle et un rectangle, et n'aurait aucune raison de croire, *a priori*, qu'il s'agit de deux perspectives portant sur un même objet de la troisième dimension.

De façon analogue, nous voyons la réalité tantôt sous le prisme de l'esprit, tantôt sous le prisme de la matière, et nous soupçonnons qu'il existe un lien entre ces deux perspectives — que nous cherchons à corps perdu dans le cerveau. Nous faisons donc l'hypothèse (1) que l'esprit et la matière sont deux réalités différentes et (2) que ces réalités présentent un point de rencontre dans le cerveau. Nous pensons en outre que (3) la réalité de l'esprit doit pouvoir s'expliquer uniquement par celle de la matière, si bien que les termes se rapportant à esprit (pensées, idées, sensations, douleur...) ne sont que des phénomènes physiologiques.

J'aimerais suggérer tout le contraire ; à savoir (1) que l'esprit et la matière sont une seule réalité, ni purement mentale, ni purement spatiale et (2) que la réalité ne peut être entièrement expliquée par l'esprit, ni entièrement expliquée par la matière. Par suite de quoi, il semblerait que notre compréhension du monde exige de passer par l'une et par l'autre de ces perspectives.

Il existe des problèmes qui reçoivent une réponse adéquate sous l'angle de l'esprit et d'autres qui exigent de prendre l'angle de vue de la matière. Il s'agit donc d'un monisme (une seule réalité existe) à double aspect (cette réalité peut prendre deux visages).

Pour rendre ma position philosophique crédible, je chercherai à défendre que le monisme à double aspect est compatible avec la pratique neuroscientifique. Je n'aurai pas l'ambition de brasser, de façon transversale, toutes les théories neuroscientifiques un tant soit peu influentes. Je jetterai plutôt mon dévolu sur la théorie de la conscience de Gerald Edelman et de Giulio Tononi. Ce choix dissimule un double parti pris. Premièrement, celui de se focaliser plus spécifiquement sur la question de la conscience et, d'autre part, celui de s'inscrire dans une théorie favorable à l'interprétation métaphysique du monisme à double aspect. Je vais préciser, et justifier autant que possible, ces deux angles d'attaque.

Concernant tout d'abord le choix d'étudier les théories neuroscientifiques de la conscience, il s'impose de façon assez pertinente lorsque l'on prend la peine de se rappeler que l'hypothèse du monisme à double aspect tente d'apporter une réponse au problème de l'articulation entre le corps et l'esprit. Quel objet d'étude, plus que la conscience, nous heurte de front avec ce problème théorique ? La conscience ne peut pas faire l'objet d'une expérience neutre, comme il l'est généralement admis de la part des objets tombant dans le giron de la démarche scientifique. Il n'est pas possible de la placer en face d'un télescope ou d'en soumettre un échantillon au microscope, puisque c'est la vie interne du sujet qu'il est question d'étudier. Bien qu'elle soit au centre de la perception — ou pour cette raison même — elle semble échapper aux lois connues de la diffraction. C'est un couteau au fil certes aiguisé, mais incapable de se couper lui-même.

Le deuxième parti pris consiste à choisir précisément la théorie d'Edelman et de Tononi, assez favorable à l'apport de la démarche philosophique, plutôt que des modèles de la conscience d'accès de Dehaene, Changeux ou Naccache¹⁹. Ce choix s'explique en raison de la méthodologie toute particulière choisie par Tononi pour approcher la conscience. Plutôt que de partir des théories existantes de l'attention ou de fonctionnement cérébral afin d'isoler un certain nombre de processus clefs, articulés par exemple autour du système thalamo-cortical, il décide de partir de l'expérience phénoménologique de conscience, du « Je » qui voit le monde, qui sent, qui pense, qui

19. Ces modèles de la conscience d'accès se fondent originellement sur l'intuition de William James d'après laquelle il existe un système nerveux supérieur qui observe et contrôle les autres systèmes cérébraux. D'après un discours de Stanislas Dehaene au Collège de France en 2009-2010, ces modèles reposent sur quatre idées, développée dès les années 50-60 et ayant reçu progressivement une plus grande adhésion : (1) la conscience est associée à un système de supervision centrale, (2) que ce système est à capacité limitée et doit opérer des choix sur l'information à traiter (système du goulot d'étranglement), (3) que ce système nécessite des boucles de rétroaction, aussi appelée réentrances et (4) la conscience fournit un espace interne de synthèse, de maintien et de partage de données.

doute à l'occasion.

Dans cette expérience originelle, la conscience se donne comme la totalité de la réalité, mais elle répond à un certain nombre de lois d'organisation. Tononi cherche à les identifier au moyen de cinq axiomes, après quoi il postule l'existence d'une analogie de structure entre le fonctionnement de la conscience phénoménologique et les principes physiologiques de traitement de l'information par l'organisme biologique. Il cherche donc, pour le dire en d'autres termes, une traduction neurologique de principes acquis premièrement par le biais d'une démarche introspective.

Cette façon de faire, chez Tononi, présente plusieurs similarités eu égard à ma thèse. Premièrement, elle accepte l'idée qu'il y ait présence, dans la théorie scientifique de la conscience, d'un point de vue en Je. Plutôt que de le laisser à l'état de supposition implicite, il cherche à en donner les lois précises ; il accepte donc l'idée que l'on puisse construire un modèle de l'esprit, ce que j'ai appelé plus haut *un modèle psychologique*.

Par ailleurs, et ce n'est pas un moindre avantage, sa démarche peut aisément s'accommoder du monisme à double aspect. Elle ne réduit en effet pas la conscience à des substrats matériels, mais les envisage comme des processus d'information. Il est donc admissible d'interpréter — bien que ce ne soit pas là son ontologie — cette information comme une réalité pouvant recevoir l'interprétation complémentaire de l'esprit et de la matière.

Je m'efforcerai donc d'exposer cette théorie de Tononi par le menu, en montrant les inspirations qu'elle a reçue de Gerald Edelman. Les deux auteurs ayant collaboré ensemble, il m'a semblé difficile de les séparer. Je tenterai d'isoler les points de rencontre et les divergences apparaissant progressivement dans leurs travaux respectifs. Mon objectif ne sera pas, simplement, de montrer que les neurosciences sont favorables au monisme à double aspect, ni de soutenir que celui-ci pourrait offrir à l'intuition empirique de nouvelles pistes de recherche. Une telle ambition serait de l'ordre d'une discussion assez superficielle de la littérature. Je dis « superficiel » au sens où cela ne justifie pas fondamentalement la pertinence de ces auteurs dans le cadre de ma thèse.

Au contraire, l'analyse de Tononi visera à poser les fondations, ou disons plutôt les réquisits indispensables, d'une théorie bi-perspectiviste de la conscience. Cette théorie doit partir du point de vue à la première personne (introspection), mais sans supposer que ce point de vue en Je soit quelque chose de donné et d'immédiat. Elle doit en outre saisir l'articulation entre l'unité du « Je pense », « Je sens », « Je suis » avec la multiplicité des objets de pensée, de sensations ou des états de conscience dont cette unité est constituée. La théorie méréologique²⁰ de la conscience que j'appelle de mes

20. La méréologie est la discipline qui étudie les relations de parties : les relations entre une partie et son tout, ainsi que les relations entre plusieurs parties au sein d'un même tout (Cotnoir et Varzi 2021).

vœux sera esquissée à la fin de ma thèse, en reprenant la théorie tricamérale de l'esprit humain, qui distinguait la raison, la perception et la sensation. Elle sera développée plus amplement dans un essai de 300 pages, intitulé « La Conscience est-elle morte ? », dont je joins l'introduction à la fin de ce document (cf. annexe 7).

Le Problème du jugement en philosophie

De l'analytique au synthétique

Chapitre 1

Le Jugement chez Platon : L'Origine du problème

*Nous ne savons pas et nous ne
saurons jamais.*

Emil du Bois-Reymond

Une question qui est souvent débattue dans l'histoire des idées, et dont la simple évocation fait trembler les savants à la barbe blanche, consiste à se demander si les problèmes philosophiques sont susceptibles de recevoir des réponses nouvelles. À bien regarder la façon dont l'édifice de la philosophie s'est érigé à travers les siècles, on peut avoir l'impression que chaque pierre ajoutée au sommet de la tour brinquebalante des concepts métaphysiques a été prélevée dans les fondations de cette tour ; si bien que la philosophie, loin de croître vers les cieux à la façon de la physique ou des mathématiques, n'a jamais cessé de stagner.

Si tel est vrai, cela signifierait que nos innovations philosophiques ne seraient rien d'autre qu'un peu de peinture, de dorure, d'inscriptions ingénieuses ajoutées sur les briques fissurées des penseurs qui nous ont précédés. Ces embellissements auraient certes le mérite d'affiner ci et là les lignes de réflexion ouvertes par nos prédécesseurs, mais leur rôle le plus important serait surtout de doter les vieilles positions philosophiques d'un langage contemporain. Ce langage consiste à donner une structure logique, une écriture mathématique, à des énoncés qui n'étaient jusqu'à présent rien que des intuitions. Les problèmes épistémologiques, qui se présentaient à l'époque de Socrate comme des guerres de définition ou de principes universaux, se trouvent changés, au sein de nos théories modernes, en équations mathématiques.

Un exemple de ce phénomène, que peu de gens, je crois, contesteraient, peut être trouvé dans le concept de l'atomisme. L'idée que le monde soit le produit des lois de la nature, entendue comme la collision des atomes, est en

substance une idée de Démocrite. À l'époque de la Grèce antique, cette idée n'était pas nécessairement majoritaire — loin s'en faut — mais les tenants et aboutissants de cette idée avaient déjà été étudiés quant à leurs termes les plus généraux, les plus abstraits, les plus conceptuels. Au rang de ceux-ci, on se demandait si le vide existait, comment le mouvement était possible ou quelle force mouvait la matière ? Un échantillon de ces réflexions peut être trouvé dans l'argument de Démocrite ci-dessous :

Si tout corps est divisible à l'infini, de deux choses l'une : ou il ne restera rien, ou il restera quelque chose. Dans le premier cas, la division ne saurait aboutir à un néant pur et simple, car alors la matière n'aurait qu'une existence virtuelle, dans le second cas on se pose la question : que reste-t-il ? La réponse la plus logique, c'est l'existence d'éléments réels, indivisibles et insécables appelés donc atomes^[1]

La science moderne, en tant que paradigme, a le mérite d'avoir repris certaines positions philosophiques et de leur avoir donné un développement systématique. Certains problèmes pratiques, comme la manière dont les atomes peuvent interagir les uns les autres, ou les propriétés résultantes de la matière, ont reçu des progrès substantiels — progrès d'autant plus impressionnants qu'ils ont abouti à des applications transfigurant le visage du monde. Les énigmes les plus fondamentales, en revanche, telles que le rapport entre le vide et la matière, ou — dans un autre registre — celui de savoir ce qu'il y a avant le Big Bang, restent largement inélucidées. Il serait illusoire de penser, sous prétexte que ces problèmes aient reçu une formulation mathématique plus précise, que nous ayons avancé significativement dans la compréhension des mystères métaphysiques.

Cette impasse vient peut-être de ce que la méthode scientifique, quoique d'une efficacité prodigieuse pour modéliser le monde, ne permet pas de répondre aux grandes questions philosophiques ? Le savoir moderne s'inspire en effet de la méthode de Descartes, d'après laquelle il faut réduire un problème en ses constituants les plus simples, « pour s'élever peu à peu, comme par degrés, jusqu'à la connaissance la plus composée »^[2]. Or de quel droit pouvons-nous être sûrs que l'Univers soit découparable en simples constituants ? Et si c'est bien le cas, le découpage est-il univoque ou s'apparente-t-il aux constructeurs de puzzle qui, partant d'une même image, et suivant certaines règles de décomposition, donnent naissance à une pléthore de puzzles distinctifs ? Est-il permis de penser que, en suivant un autre schéma méthodologique, ce soit non pas la matière, mais bien le vide, la conscience ou la volonté qui constituerait la cause au fondement de toute « réalité » ? (ce mot

1. Cet argument de Démocrite est rapporté par Aristote dans *De la Génération et de la corruption*, livre I, chap. 2, 9 à 19.

2. DESCARTES, *Discours de la méthode*, troisième règle.

étant peut-être, comme le disait si bien Nabokov, le seul qu'il est légitime de mettre entre guillemets³).

La science n'a pas la même visée que la philosophie. Alors que cette dernière part des choses les plus simples pour se hisser peu à peu à des conceptions « si paradoxales que personne ne veut y croire » (dixit Russell), la science, au contraire, part de la complexité du monde et la reconduit, *manu militari*, à un ensemble de principes les plus simples qu'il soit permis de produire. Elle ne rechigne jamais à user en chemin, et à abuser quelques fois, du rasoir d'Ockham : *pluralitas non est ponenda sine necessitate* (les multiplications [d'un même élément explicatif] ne doivent pas être utilisées sans nécessité). Son but est de produire, non pas de la complexité, mais de la confiance dans le monde.

C'est pourquoi je ferai l'hypothèse, dans cette thèse, que certains problèmes techniques auxquels nous heurtons dans différentes disciplines scientifiques — en ce compris l'intelligence artificielle — sont avant tout des problèmes de concepts. Ce que j'entends par là, c'est que les difficultés qui s'y rencontrent ne sont pas *uniquement* du ressort de l'observation empirique du monde, pas plus que de la résolution analytique d'une équation, mais qu'elles dépendent *aussi* de la façon dont nous réduisons un problème à des unités d'analyse minimales. Des questions telles que celles-ci : « qu'est-ce que la pensée ? », « une machine peut-elle juger du vrai et du faux ? » ne recevront pas seulement une réponse en améliorant la puissance de nos ordinateurs ou notre compréhension du cerveau humain. Elles dépendent également des catégories que nous utilisons pour décrire les phénomènes se produisant dans notre conscience. Les « idées », les « données des sens », les « sensations », les « perceptions », la « mémoire », les « raisonnements » sont des abstractions posées sur un tout cognitif complexe. Il y a dans le choix des concepts que nous manipulons, et dans la façon dont nous les modélisons avec nos ordinateurs, un caractère potentiellement arbitraire et réducteur.

À ce titre, les méthodes d'intelligence artificielle ne font pas exception. À l'heure où les réseaux de neurones semblent, plus que jamais, être en passe de dépasser ou d'imiter l'être humain sur toutes sortes de tâches cognitives autrefois irréalisables — en ce compris la création d'images, la composition de musique et l'usage du langage — il est peut-être temps de se demander quelles réductions ont été réalisées dans le processus ? Qu'a-t-on perdu de la richesse de la conscience humaine en la faisant entrer, *manu militari*, dans l'avaloir à big data que sont devenus les algorithmes contemporains ? Le jugement des machines s'appuie-t-il sur les mêmes concepts que le jugement humain ? Et que peut-on attendre, ultimement, de l'intelligence artificielle ?

Hypothèse 2 (Hypothèse de recherche). *Les obstacles auxquels se heurte le développement de l'intelligence artificielle ne relèvent pas uniquement de*

3. Je fais référence à une simple parenthèse que l'auteur susnommé introduit dans la postface de *Lolita*, qui traite d'un tout autre sujet.

difficultés techniques, mais dissimulent également des problématiques épistémologiques, présentes depuis les origines de la philosophie, et liées aux concepts qui sont manipulés.

Un problème de concepts qui intéresse au premier chef le développement de l'intelligence artificielle consiste précisément à savoir comment naissent les idées dans l'esprit humain. Nous faisons toutes et tous l'expérience de voir, dans l'espace qui nous entoure, un monde peuplé d'objets : une table s'étend face à nous, dotée de quatre pieds, sur laquelle s'étalent des livres, des verres, des biscuits triangulaires. Tous ces corps peuvent être abstraits un à un dans notre conscience, et détachés pour ainsi dire du contexte dans lequel nous les avons aperçus. Nous pouvons du reste imaginer ces objets même en l'absence d'un exemplaire concret, en nous représentant dans l'esprit tantôt une forme, tantôt une couleur, tantôt une sensation générale.

Il n'est toutefois pas aisé de comprendre comment se forme cette abstraction qu'on appelle une idée. De quel droit la vision d'un triangle, par exemple, se rapporte-t-elle au concept du triangle ? Le concept du triangle n'est pas ce triangle en particulier, mais il englobe pourtant l'ensemble des triangles possibles et imaginables. Chaque fois que je me représente un triangle, en déployant pour ce faire toutes les forces de mon imagination, ce n'est pas le concept de triangle que j'ai sous les yeux, mais tel ou tel triangle particulier, doté d'une forme, de certains angles, d'une couleur, d'un parfum pour ainsi dire. Sans le concept de triangle, qui semble décidément échapper à toute représentation, serais-je capable d'identifier les triangles dans la nature ? Pourrais-je les isoler les uns des autres ? Les manipuler par la seule inspection de mon esprit ? Trouver des propriétés concernant la somme de leurs angles ?

Dans les chapitres qui suivent, je souhaiterais montrer comment ces problèmes se présentent sous la plume de Platon. Ce choix de vous parler du père de la philosophie grecque pourrait trahir un manque d'originalité. Je n'aimerais pas paraître pour l'un de ces radoteurs qui se contentent, en guise de pensée, d'enserrer comme le lierre le tronc des penseurs millénaires. J'ai toutefois deux raisons de vous parler des dialogues socratiques.

La première, et non des moindres, c'est que je souhaite montrer que le problème épistémologique auquel se heurte l'intelligence artificielle n'est pas un problème technique. Cette affirmation signifie que le problème peut être explicité sans passer par le formalisme des mathématiques actuelles, mais qu'il était déjà présent dans les origines de la philosophie. Tous les problèmes véritablement fondamentaux, en ce qu'ils touchent à la nature même de la pensée, ont été traités par les premiers êtres humains capables d'objectiver la pensée et de la mettre par écrit. Or les plus vieux écrits de philosophie parmi ceux à avoir été conservés nous viennent précisément de la Grèce antique.

La seconde raison, plus importante encore, c'est que Platon construit

des concepts qui ne sont pas sans intérêt pour éclaircir certaines difficultés rencontrées dans la pratique actuelle de l'intelligence artificielle. Un des mérites du maître de l'Académie d'Athènes consiste en effet à avoir distingué, au sein de la pensée humaine, trois espèces de jugements : les jugements d'expérience, les jugements vrais et les jugements vrais munis de justification.

Les jugements d'expérience, tout d'abord, peuvent être vus comme une certaine capacité de synthèse ; ils se rapportent au monde réel en procédant à des observations, des comparaisons, des imitations. Les jugements vrais munis de raison, quant à eux, reposent sur une aptitude au raisonnement logique. Ils partent de certains principes et s'élèvent vers des conclusions plus générales. Les jugements vrais sans justification, pour finir, occupent un point médiant. J'aimerais montrer qu'ils représentent peut-être, en philosophie, le prototype le plus archaïque de cette espèce de jugement proprement intuitif dont j'ai parlé dans l'introduction (voir [1]). Ces jugements ne sont ni purement empiriques, ni purement logiques. Ils assurent une liaison entre la sensation et la raison, entre l'expérience et la déduction.

1.1 Socrate vs Protagoras : Les Monologues de l'âme

Pour éclaircir le rôle précis que Platon confère au jugement, la lecture d'un texte intitulé le *Théétète* présente plusieurs intérêts⁴. Quant à sa méthode, tout d'abord, ce texte fondateur de Platon nous introduit à l'usage du logos dans l'investigation philosophique ; quant à son objet, il pose la question de savoir ce qui rend possible une science du changement ; quant à sa visée polémique, pour finir, il est en étroite dialogue avec les sophistes et les présocratiques. C'est pourquoi j'aimerais m'appuyer sur cette œuvre pour suggérer en quoi Platon se sépare de la tradition antérieure, celle des sophistes, et conduit du même coup la philosophie à un genre spécifique de jugement : le jugement synthétique *noétique*.

Le *Théétète*, comme de nombreux dialogues socratiques, vise à découvrir, au travers d'un échange entre plusieurs personnages, l'essence d'un concept. Le concept dont il est question ici, c'est la science, entendue comme la connaissance vraie. On comprend que ce thème est d'une importance toute particulière pour la question qui nous intéresse, puisqu'il est question de déterminer comment l'esprit humain peut déterminer la différence entre le vrai et le faux.

Tout commence tandis que Socrate, en visite chez son ami Théodore, demande à ce dernier de lui présenter le plus beau et le plus talentueux de ses disciples. Sur ce, Théodore l'introduit auprès du jeune Théétète, beau

4. Dans la suite du texte, je citerai les passages du *Théétète* en utilisant le système de référencement Estienne. La traduction française utilisée est celle de CHAMBRY.

quant à son âme. Il fait de lui le portrait du parfait philosophe : « Je n'ai encore vu personne si admirablement doué. Qu'un homme ait une facilité à apprendre telle qu'on trouverait difficilement le pareil, qu'il soit doux au suprême degré et qu'il ait avec cela un courage sans égal, je n'aurais jamais cru que cela pût se rencontrer et je n'en vois point d'autre exemple⁵ » Socrate, enchanté, entreprend d'interroger le jeune garçon. Il lui demande si la science est sagesse, ou s'il existe au contraire une différence particulière entre ces deux types de vertus.

Les premières réponses du jeune homme à Socrate, on s'en doute, ne sont pas satisfaisantes. À la question de savoir ce qu'est la science, il en énumère les différentes disciplines : la géométrie, l'astronomie, l'arithmétique et la musique sont des sciences, dit-il, mais aussi la cordonnerie et les différents arts des ouvriers. Ce faisant, il n'en montre pas l'essence. C'est pourquoi Socrate a beau jeu de le reprendre : « Ce n'est pas la science que tu définis là, mais ses objets. Si je te demandais ce qu'est la boue et que tu me répondes : Il y a la boue des potiers, celle des briquetiers et d'autres, je ne serais pas du tout renseigné sur la nature de la boue. Ce qu'il fallait me dire, c'est que la boue est de la terre délayée dans de l'eau⁶ »

Socrate précise d'emblée qu'il ne détient pas la réponse à leur recherche, mais qu'il est en mesure d'évaluer les idées originales que Théétète lui donnera. Il est en sus capable de stimuler son intelligence en examinant avec lui les théories de ceux qui se sont déjà exprimés sur la question. À ce titre, il opère un peu à la manière d'un accoucheur des âmes : « ... n'as-tu pas entendu dire que je suis le fils d'une sage-femme, et que j'ai l'art d'accoucher les esprits, comme elle avait celui d'accoucher les femmes ? Je sais de plus discerner si l'esprit d'un jeune homme enfante une chimère ou un fruit réel et vrai. Fie-toi donc à moi et ne te chagrine pas si, en examinant ce que tu diras, je juge que ce n'est qu'un fantôme sans réalité⁷ »

Ils vont alors envisager tour à tour trois définitions. La première, qui est aussi celle sur laquelle Platon s'étend le plus, consiste à dire que « la science n'est autre chose que la sensation⁸ » C'est en fait l'opinion de Protagoras. La seconde, plus subtile, consiste à dire que « ce n'est donc point dans les impressions que réside la science, mais dans le raisonnement sur les impressions⁹, ce que Théétète formule en disant que la science est l'opinion vraie. La troisième, enfin, est une adaptation de la précédente, et consiste à dire que la science est l'opinion vraie munie de raison, c'est-à-dire, munie d'un ensemble de définitions plus analytiques des termes impliqués dans l'opinion. Ces trois approches feront chacune l'objet d'apories. Leur analyse donne cependant à Platon l'occasion d'exposer et de critiquer les

5. *Théétète*, 144a.

6. cf. introduction au *Théétète* par CHAMBRY, p. 44.

7. *Ibidem*, p. 45.

8. *Théétète*, 151e.

9. *Théétète* », 186d.

thèses en vigueur à son époque, notamment celle de Protagoras.

1.1.1 Le Jugement d'expérience

Il y a une raison pour laquelle Platon accorde tant d'importance à la première thèse, celle de la sensation. C'est qu'elle semble avoir été celle de Protagoras, un de ses plus importants rivaux. La thèse de Protagoras est la suivante : la science est sensation et la sensation est changeante. En guise d'illustration, il donne l'exemple du chaud et du froid. Un même vent qui sera cause, pour l'un, de frissons, pourra sembler tiède à un autre. De sorte que ce qui doit être tenu pour vrai, ce n'est pas un quelconque royaume d'objets auquel nous accéderions par l'entremise de nos sens, mais ce sont les sensations subjectives, et toutes relatives, que nous éprouvons. Le jugement de la science est ainsi conçu comme un *jugement d'expérience*.

Socrate avance une succession d'objections à l'encontre de cette thèse de Protagoras. La plus célèbre consiste à accepter la thèse et à montrer qu'elle conduit à une absurdité. Elle fonctionne de la façon suivante : supposons que Protagoras tienne pour vraie la proposition P selon laquelle « l'Homme est la mesure de toutes choses ». Dans le même temps, supposons que Socrate nie cette proposition. Du propre point de vue de Protagoras, il faudra que la proposition soit à la fois vraie (puisque les choses sont à la mesure de Protagoras) et à la fois fausse (puisqu'elles sont aussi à la mesure de Socrate). La formalisation de ce raisonnement n'est pas sans rappeler un syllogisme célèbre :

Protagoras croit que l'Homme est la mesure de toutes choses.

Mais Socrate est un Homme.

Donc Protagoras croit que Socrate est la mesure de toutes choses.

Or comme Socrate tient P pour fausse, on arrive à la conclusion paradoxale que Protagoras ne croit pas que l'Homme est la mesure de toutes choses. Notre brave sophiste serait-il pris au piège ? Socrate conclure, ironiquement : « Supposons qu'il [Protagoras] n'ait pas cru lui-même que l'Homme est la mesure de toutes choses, et que le grand nombre ne le croie pas non plus, comme, en effet, il ne le croit pas, ne serait-ce pas alors une nécessité que la vérité telle qu'il l'a définie n'existât pour personne^[10] ? »

1.1.2 Le Jugement vrai sans justification

Après avoir sévèrement contesté le sensualisme, Socrate finit par convaincre Théétète d'en venir à une nouvelle définition. Si la connaissance ne provient pas des impressions sensibles, elle doit provenir du raisonnement sur ces impressions, c'est-à-dire du *jugement*. L'introduction de ce dernier mot doit

10. *Théétète*, 170e-171a.

être, je crois, soulignée, car il indique un mouvement d'importance majeure dans l'histoire de la pensée. Il pose, conjointement avec la capacité de sentir, une autre espèce de faculté. Mais qu'est-ce que « juger » dans la bouche de Socrate ? C'est « le nom [...] qu'on donne à l'âme lorsqu'elle s'applique elle-même toute seule à l'étude des êtres¹¹. » Selon Socrate, le jugement est ainsi cette capacité qu'a l'âme à converser avec elle-même, dans un mode dialogique, afin de découvrir les êtres dont répondent les sensations. Il est probable que le type de dialogue auquel l'âme se livre est analogue à celui qui s'est construit entre Socrate et Théétète.

Bien sûr, la science ne se compose pas indifféremment de l'ensemble de tous les jugements possibles, mais seulement de ceux qui sont vrais. Théétète avance ainsi la suggestion suivante : « Dire que toute opinion est science, Socrate, c'est impossible, puisqu'il y a aussi une opinion fausse. Mais il y a des chances que l'opinion vraie soit science, et voilà ma réponse¹². » Socrate lui demande alors de dire ce qui fait d'une opinion qu'elle est fausse. Il va même plus loin : il demande ce qui rend possibles les faux jugements. En d'autres termes, il demande à Théétète de lui fournir un modèle de la connaissance qui puisse rendre compte de l'erreur et de l'ignorance.

Socrate entreprend alors de dessiner avec Théétète les contours d'une théorie de la connaissance. Pour ce faire, il recourt à une allégorie, celle du morceau de cire : « Suppose donc avec moi, pour le besoin de l'argument, qu'il y a dans nos âmes un bloc de cire¹³. » Chaque fois que l'esprit veut se souvenir de quelque chose qu'il a vu, entendu ou conçu, l'empreinte de cette chose est appliquée sur le bloc de cire. Sous cette hypothèse, se souvenir consistera à relever l'empreinte qui est dans notre âme et à l'appliquer à la chose sensible appropriée. Si par exemple l'on a la marque de Théodore et de Théétète, il s'agira de retrouver ce souvenir dans notre âme et de le faire correspondre adéquatement aux personnes que l'on a devant les yeux. Mais lorsque la marque n'a pas été réalisée ou qu'elle s'est effacée, alors il est impossible de se remémorer correctement de la chose. L'erreur résidera dans une mauvaise correspondance entre les marques et les objets sensibles.

Aussi convaincante qu'elle puisse paraître de prime abord, cette théorie ne manque pas d'éveiller le scepticisme aiguisé de Socrate. L'allégorie du bloc de cire semble laisser entendre que la connaissance est toujours une affaire d'adéquation entre le jugement et la sensation. Or n'est-il pas possible de raisonner dans l'intelligible ? L'erreur n'est-elle pas aussi possible en l'absence d'objets sensibles ? Il suffit de penser aux démonstrations mathématiques pour se rendre compte que les notions de vrai et de faux ne sauraient, sauf à défendre une conception strictement empirique des mathématiques, se limiter à une correspondance entre la pensée et la sensation.

11. *Théétète*, 187a.

12. *Théétète*, 187b.

13. *Théétète*, 191c.

L'exemple qu'il donne est celui du calcul mental. N'est-il jamais arrivé qu'un homme, considérant les chiffres 7 et 5, se mette à croire que leur somme vaut 11 ? « Il faut donc montrer que l'opinion fausse est tout autre chose qu'un désaccord de la pensée et de la sensation. Si en effet c'était cela, nous ne nous tromperions jamais dans nos pensées pures^[14]. »

Socrate se propose alors d'élucider la signification des termes « connaître » et « savoir ». Pour commencer, il opère une distinction entre « avoir la science » et la « posséder ». Au même titre que l'on peut posséder un manteau sans l'avoir présentement sur soi, l'on peut être détenteur d'un savoir sans l'avoir présentement à l'esprit. Dès lors que la pensée est mobilisée, dès lors que la mémoire est secouée, les idées concrètes correspondant à ce savoir font surface dans l'âme. C'est comme si chaque savoir détenu était l'oiseau d'un colombier. Lorsque l'esprit souhaite invoquer ce savoir et l'utiliser dans un raisonnement, il doit rechercher le volatile dans son enclos, ce qui demande un certain effort. Le travail est certes moindre que pour attraper l'oiseau à l'état sauvage ; il est plus facile d'appliquer que d'apprendre pour la première fois. N'empêche que l'erreur reste possible. Le risque est toujours présent de prendre un oiseau pour un autre, ou de ne pas parvenir à attraper l'oiseau correspondant au savoir mobilisé.

Ceci explique deux choses : d'une part, la possibilité des erreurs de raisonnement, comme les fautes de calcul. D'autre part, le fait que même un arithméticien accompli, qui connaît tous les nombres, puisse continuer d'en chercher leurs propriétés. Posséder le savoir des nombres, ce n'est pas avoir les nombres, et leurs caractéristiques, présentement devant les yeux. Nous reviendrons plus en détail sur cette théorie lorsque nous la mettrons en comparaison avec la conception kantienne.

1.1.3 Le Jugement vrai muni de raison

Un doute persiste toujours et s'immisce, meurtrier, dans l'esprit le plus tenace. Dira-t-on d'un homme qu'il est savant lorsqu'une opinion vraie se glisse accidentellement entre ses lèvres ? Ou attendons-nous de lui qu'il puisse fournir la justification de sa croyance ? Socrate prend l'exemple d'un juge qui, pour trancher sur une matière, a à sa disposition un ensemble de témoins oculaires. Il ne connaît pas la situation à titre personnel, car il ne l'a pas vécue et n'y a pas assisté lui-même. De son propre aveu, le savoir n'est pas de son côté, mais de celui de la victime, du criminel et de ceux qui se trouvaient là. Il se peut malgré tout qu'il prenne la bonne décision. « Si l'opinion vraie chez les juges et la science étaient la même chose, jamais le meilleur des juges n'aurait une opinion droite sans science. Mais en fait il apparaît que ce sont deux choses différentes^[15]. »

14. *Théétète*, 196c.

15. *Théétète*, 164b.

Cela suggère à Théétète une troisième et dernière définition : « l'opinion vraie accompagnée de raison est science, mais [...] dépourvue de raison, elle est en dehors de la science¹⁶ » Que faut-il entendre ici exactement par « raison » ? Des définitions ? Des raisonnements ? Socrate expose une théorie. L'esprit, dit-il, ne peut rien connaître des premiers éléments dont est constituée toute réalité. Il ne peut, tout au plus, que les nommer, non les décrire : « Chaque élément pris en soi ne peut que se nommer, et l'on n'en peut rien dire de plus, ni qu'il est, ni qu'il n'est pas ... il ne faut rien lui accoler, si l'on veut exprimer cet élément seul. On ne doit pas même y joindre ces mots : « le », ni « cela », ni « chacun », ni « seul », ni « ceci », ni beaucoup d'autres mots semblables¹⁷ » En extrapolant légèrement les propos de Socrate, peut-être n'est-il pas trop audacieux de dire la chose suivante : la connaissance d'un objet consistera à joindre l'intuition correcte de l'âme (qui nomme l'objet et est de l'ordre du synthétique) à une explication rationnelle correcte (qui décompose l'objet en ses éléments est de l'ordre de l'analytique).

Je souhaite attirer l'attention sur le fait que la structure des idées dans le *Théétète* n'est pas laissée au hasard, mais témoigne au contraire d'une certaine anthropologie. Platon commence d'abord avec ce qui relève de la sensation, caractérisée par le changement. La question qu'il pose est celle de savoir quel est le rapport de l'organe sentant à l'objet sensible. Au fil de la conversation, il pose la question : les sons et les couleurs sont perçus grâce aux oreilles et aux yeux, mais le sont-ils "par" eux ou "avec" eux ? Précisons bien cette nuance. Que les objets sensibles soient perçus *avec* les organes signifie que le sujet est muni de facultés sensibles qui sont comme greffées sur son corps et qui les lui font connaître. Mais s'ils sont connus *par* les organes, cela laisse entendre que c'est une même faculté de l'âme qui s'exerce au travers des organes et qui, par leur entremise, perçoit les objets sensibles. Ces deux approches sont radicalement différentes. Dans le premier cas, les sens se suffisent à eux-mêmes, et il n'est pas besoin d'ajouter un autre ordre d'intelligibilité. Dans le second cas, ils ne sont que les agents d'une faculté supérieure.

Pour prouver l'existence de cette faculté, Platon entreprend de montrer que certaines idées, celles se rapportant à "l'être" et au "non-être", ne peuvent pas provenir des sens. Considérons un son et une couleur. Il n'est pas possible de connaître la couleur au moyen de l'oreille, pas plus qu'il n'est possible de connaître le son au moyen de la vue. Cependant, il est possible d'avoir du son et de la couleur des idées se rapportant aussi bien à l'un et à l'autre. C'est le cas, notamment, de l'idée que le son et la couleur existent tous deux, ou encore, de l'idée de leurs ressemblances et de leurs différences. C'est le cas, en bref, de tout ce qui se rapporte à leur être. Socrate demande alors :

16. *Théétète*, 201c-201d.

17. *Théétète*, 202a.

« par quel organe conçois-tu tout cela de ces deux sens ? Ce n'est ni par l'ouïe ni par la vue qu'il est possible de saisir ce qu'ils ont de commun¹⁸. » Il faut donc qu'il existe une faculté qui surplombe les sens. Cette faculté est celle qui use de la langue. Elle ne dépend d'aucun organe, mais est exercée par l'âme elle-même. C'est elle qui traite « de l'être et du non-être, de la ressemblance et de la dissemblance, de l'identité et de la différence, et aussi de l'unité et des autres nombres appliqués à ces choses¹⁹. »

Il faut donc opérer une distinction entre les impressions sensibles comme le « dur », le « mou », le « froid » le « chaud », qui sont données par des organes corporels, et les idées portant sur ces impressions, qui eux sont des jugements de l'âme. Ces dernières seules touchent à l'être des choses et à leur vérité. « Il y a donc certaines choses que, dès leur naissance, les hommes et les bêtes sont naturellement capables de sentir : ce sont les impressions qui gagnent l'âme en passant par le corps. Au contraire, les raisonnements qu'on fait sur ces impressions, relativement à leur essence et à leur utilité, ne viennent que difficilement et à la longue, à force de travaux et d'étude, à ceux chez qui ils se forment²⁰. »

Loin de s'arrêter à la faculté de la "sensation", qui nous donne à voir le monde par l'entremise des sens, Platon insiste sur l'existence d'une faculté de l'âme qui est purement intelligible. Il nomme cette faculté la raison (λογος). Ce terme est introduit pour la première fois lorsque Socrate et Théétète se rendent compte que juger, ce n'est pas seulement faire correspondre des idées à des impressions, mais qu'il y a aussi une forme de pensée pure, ne se rapportant à aucun contenu sensible. Ce genre d'opération est, notamment, le calcul mental. « Il faut donc montrer que l'opinion fausse est tout autre chose qu'un désaccord de la pensée et de la sensation. Si en effet c'était cela, nous ne nous tromperions jamais dans nos pensées pures. Mais, en réalité, ou il n'y a pas d'opinion fausse, ou il est impossible qu'on ne sache pas ce qu'on sait²¹. » Tant que les idées se rapportent à des choses sensibles, le vrai et le faux proviennent de leur correcte adéquation. Mais dès lors que l'on s'élève dans la pensée pure, l'erreur ne peut venir, si elle existe, que du bon ou du mauvais usage de la raison.

La nature précise de la raison en tant que faculté n'est cependant pas thématifiée dans le *Théétète*. Bien au contraire, elle peut sembler, à bien des égards, n'être qu'une opération de langage. L'on a le sentiment que la fin du dialogue peine à clarifier le concept de raison. Les définitions qui sont envisagées, puis tour à tour réfutées, sont probablement une compilation des idées soutenues par les contemporains de Platon, comme Antisthène ou l'école des Mégariques. Quelles que soient leurs origines, la multiplicité de ces théories montre l'importance que cette question revêtait à l'époque dans les

18. *Théétète*, 185b.

19. *Théétète*, 185b-185c.

20. *Théétète*, 186b-186c.

21. *Théétète*, 196c.

cercles philosophiques, et renforce du même coup le caractère inaugural de la réponse apportée par Platon lorsqu'il choisit de faire du concept naissant de la Raison le point d'appui de toute la science. Il reste à déterminer comment exactement il définit cette faculté et quelle place il lui donne dans sa théorie du jugement.

1.2 La Contemplation du Beau

J'ai indiqué que Platon envisageait trois types de jugements : le jugement d'expérience, le jugement vrai sans justification et le jugement vrai muni de raison. Le premier se fonde sur l'observation du monde et représente le versant proprement synthétique de la connaissance. Le dernier se fonde sur le raisonnement et représente, quant à lui, le versant analytique. Entre ces deux falaises idéales, pétrées d'ambitions et de rêves déçus, se tient un jugement qui est vrai, mais non justifié. Platon ne reconnaît cependant pas, à ce jugement, plus de valeur qu'un simple jugement d'expérience. Il n'envisage pas l'existence d'une voie médiane entre la synthèse et l'analyse. Dans la bouche de Socrate, il n'y a pas de tiers exclu : soit une chose relève des sens, et est potentiellement trompeuse, soit elle relève du raisonnement.

Cette dichotomie entre les sens et la raison est peut-être au fondement de la conception occidentale de la rationalité. La vocation philosophique chez les anciens Grecs, puis tout au long du Moyen Âge, ne relève-t-elle pas d'une suspicion généralisée à l'égard des sens ? À l'exception des empiristes anglais, et des atomistes précoces, la vérité est le plus souvent placée du côté de l'analyse. Ce préjugé disqualifie les jugements qui ne sont pas justifiables ni argumentables. On peut comprendre que de tels jugements, en effet, pourraient difficilement participer d'une vie rationnelle collective, où la vérité fait l'objet de partages de points de vue, de débats, de délibérations. Il est sans doute difficile de séparer la vision de la vérité d'un certain projet social et politique démocratique.

Toujours est-il que cette vision binaire du jugement, telle qu'on la trouve chez Platon, s'accompagne d'une représentation bicamérale de l'esprit humain. L'âme, dans la bouche de Socrate, possède deux facultés : la sensibilité et la raison. La première s'occupe des jugements d'expérience ; la seconde s'occupe de l'analyse logique. Ces deux facultés, à leur tour, sont divisées en tout autant de sous-facultés. La sensibilité présente une part d'imagination (*ekasia*) et de croyance (*pistis*), tandis que la raison comporte une dimension calculante (*dianoia*) et contemplative (*noësis*). Cette structure de la connaissance humaine trouve une illustration éloquente dans un des livres les plus importants de Platon : la *République*. J'aimerais exposer brièvement cette théorie de Platon, car elle sera utile dans la suite de ma thèse.

1.2.1 L'Analogie de la ligne

Le livre VI de la *République*, peut-être l'un des plus importants de l'œuvre de Platon, s'ouvre sur un dialogue entre Socrate et Glaucon sur la nature du vrai philosophe — et sur le rôle de ce dernier dans la cité. À un moment donné, lassés de l'entendre réfuter les théories des autres, Glaucon et plusieurs auditeurs objurguent Socrate de leur parler de ses propres conceptions. Ils lui demandent de leur révéler en quoi le Bien consiste, de leur faire voir sa définition avec au moins autant de clarté qu'en ce qui concerne la question de la justice, traitée peu de temps auparavant.

Face à ces attentes des plus exigeantes, Socrate prévient qu'il ne pourra pas dire le Bien, qui est de l'ordre de l'intelligible, mais qu'il pourra s'efforcer de leur présenter l'une de ses métaphores sensibles : le soleil. En quel sens exactement prétend-il que le Soleil doit être apparenté au Bien ? En ce sens que la lumière est le médium indispensable à la vision. Expliquons-nous. Lorsque l'œil se tourne vers les objets que le soleil illumine, il peut les voir en toute clarté. De la même manière, lorsque l'âme se tourne vers les idées que le Bien illumine, elles lui apparaissent dans toute leur étendue. « Ce qu'il est lui [Le Bien], dans le lieu intelligible par rapport à l'intellect et aux intelligibles, celui-ci [Le soleil], l'est dans le lieu visible par rapport à la vue et aux choses visibles [...] Lorsqu'elle [l'âme] se tourne vers ce que la vérité et l'être illuminent, alors elle le pense, elle le connaît et elle semble posséder l'intellect²² »

Socrate poursuit l'analogie de façon plus spécifique. Le soleil, dit-il, est cause de la lumière et de la vue, au même titre que l'Idée du Bien est cause de la vérité et de la connaissance. Savoir et vérité sont donc des produits du Bien et, en tant que tels, ils doivent faire partie de sa définition. Cependant, le Bien ne s'y réduit pas. Il est, par sa nature, quelque chose de plus précieux encore que toutes les Idées que l'on peut s'en faire. Et pour cause : il est à l'origine de toutes les Formes. Socrate dit de lui qu'il « est au-delà de l'être (*epekeina tes ousias*)²³ », achevant ainsi d'en décrire la parfaite transcendance.

Socrate, fort de son succès, entreprend de dessiner une ligne sur le sable et la découper en quatre segments de même longueur. Les deux premiers segments, explique-t-il, représentent les deux niveaux de réalité du monde intelligible : celui des raisonnements mathématiques et des Idées pures du Bien, du Beau, du Vrai. Les deux derniers segments, quant à eux, désignent les deux niveaux de réalité du monde sensible : celui de l'imagination trompeuse et celui des croyances fondées empiriquement. La connaissance que nous avons du sensible relève de l'opinion (*doxa*), tandis que la connaissance que nous avons de l'intelligible dépend tout entière de la science (*épistémé*). Les opinions sont changeantes et capricieuses. Elles évoluent au gré de nos

22. *République*, 508b-508c.

23. *République*, 509a.

palinodies. La science, quant à elle, est détachée de l'expérience et absorbée par la contemplation des idées, en un sens qui précise et complète celui du *Théétète*.

1.2.2 Les Pathèmata

Chaque segment de la ligne est associé à un *pathèma*, une « affection ». Ce mot grec, dérivé du verbe *paschein*, veut dire « subir », par opposition à *poiein* (« faire, agir ») ou à *prattein* (« accomplir, agir »)²⁴. Il indique dans le cas présent une disposition du sujet à éprouver quelque chose au contact d'un objet sensible ou intelligible. Les quatre pathèmata définissent donc autant de « facultés » sensibles ou intellectuelles. Il existe donc deux facultés, la raison et la sensation, qui se subdivisent à leur tour en deux sous-facultés.

Noèton <i>Epistémé</i> (science)		Horatón <i>Doxa</i> (opinion)	
<i>Noësis</i> (intuitif)	<i>Diánoia</i> (discursif)	<i>Pistis</i> (croyance)	<i>Eikasia</i> (imagination)
↓	↓	↓	↓
porte sur l'anhypothétique	porte sur l'hypothétique	porte sur les objets sensibles	porte sur leurs images

L'*Eikasia* est un terme très rare dans les textes de Platon. Il n'est attesté, en tout et pour tout, que cinq fois dans le corpus de textes de Perseus²⁵, dont trois fois rien que dans la *République*. Pour se faire une meilleure idée de son usage, il est commode de se référer au verbe *eikazein* (« représenter, comparer, conjecturer »), dont il est le substantif d'action, et qui lui est beaucoup plus fréquent. Selon les commentateurs, *Eikasia* est traduit par « simulation », « imagination », « illusion », « faculté de se fonder sur des ressemblances » ou encore « représentation ». La définition de Pachet me semble la plus pertinente, dans le contexte qui nous préoccupe, car ce que Platon semble avoir bien en vue, c'est d'une capacité de juger sur les choses sensibles à partir des apparences. C'est donc la forme la plus "dévoyée" de la connaissance.

La *pistis* est la « foi », « créance », « conviction » ou encore « croyance ». Elle exprime l'idée d'une certaine attitude de confiance vis-à-vis du témoignage des sens, non seulement de la vue, mais aussi de l'ouïe, du toucher, de l'odorat, etc. C'est l'attitude consistant à croire qu'une réalité sensible se tient en face de nous, et qu'elle est telle que nous pouvons nous la représenter. En ce sens, il n'est peut-être pas trop audacieux de la rapprocher de ce que Husserl qualifiera « d'attitude naturelle ». Loin cependant de s'arrêter

24. cf. annotations de la *République* par SUZANNE [?], note 17.

25. Perseus est un projet de bibliothèque numérique initié et mené par l'université Tufts : <http://www.perseus.tufts.edu/>

aux premières impressions, c'est une attitude qui cherche une vérification, une objectivité, en interrogeant par l'expérience les données sensibles. C'est donc l'esprit de la "science" à son stade le plus empirique, le plus pratique, qui n'est pas sans évoquer les affirmations de Lachès à propos du courage. Dans un langage plus kantien, nous dirons de cette *doxa* qu'elle est formée de jugements synthétiques *a posteriori*.

La *Dianoia* est traduite par « connaissance discursive », « discursion », « raison » ou simplement « pensée ». La signification de ce terme recouvre en grande partie celui de *noësis*, que nous verrons ci-dessous, mais diffère de lui par l'adjonction de la particule « dia ». Quelle est sa signification ? Dans les usages attestés, elle peut induire une idée (1) de séparation, au sens d'une distribution ou d'une division ; (2) de pénétration ; (3) de supériorité ou encore (4) d'achèvement²⁶. Les acceptions les plus pertinentes ici sont probablement celles de la séparation ou de la division.

La faculté que Socrate a en vue, concernant ce segment, est proprement discursive. Il s'agit de structurer le contenu de la sensibilité de façon à en extraire des principes intelligibles. À la façon du géomètre qui décompose les figures géométriques en points et en segments, il est tout d'abord requis de diviser les images sensibles ou les mots du langage en autant d'éléments d'analyse que nécessaire. En posant un certain nombre d'hypothèses, l'on s'élève ensuite vers certains principes, ou certaines thèses, de validité universelle. L'archétype de ce raisonnement est celui des mathématiques, dans leur dimension formelle et apodictique. On peut penser, à titre d'exemple, aux démonstrations *ad géométrico* d'Euclide, concernant le fait que la somme des angles d'un triangle fait 180 degrés. Le discours associé est logico-déductif.

Notons, pour compléter cette définition, que la particule « dia », figurant dans le terme *dianoia*, peut aussi se référer au fait que la *dianoia* occupe l'entre-deux sur la droite de l'être. D'un côté en effet se tient l'intelligible pur, de l'autre côté se tient les images. Elle peut donc être conçue comme la faculté qui sert de médiation entre un savoir des images sensibles et un savoir des Formes. Elle s'exerce de plus au travers d'un dia-logue que l'âme se tient à elle-même.

La *noësis*, quant à elle, est traduite par « intelligence », « intellection » ou encore « activité spirituelle ». Elle est formé du verbe *noein* au moyen du suffixe *-sis*²⁷. D'usage plus rare que le mot grec *nous*, souvent utilisé pour désigner l'esprit ou l'intellect, la *noësis* a cet avantage de désigner l'intelligence en tant que faculté, et non dans ses autres sens. C'est donc bien d'une certaine capacité à appréhender l'intelligible dont il est question ici. Contrairement à la *dianoia*, qui est discursive, *noësis* ne consiste pas tant à partir d'hypothèses et à procéder par voie déductive, qu'à questionner le

26. *Ibidem*.

27. *Ibidem*.

statut de ces hypothèses, d'un point de vue épistémologique (pourquoi sont-elles vraies?) et ontologique (correspondent-elles à une réalité existant en dehors de la connaissance que nous en avons?). Il s'agit de se mettre en quête du premier principe, anhypothétique, grâce auquel il est permis de concevoir la véritable nature des choses.

Ce premier principe est celui Bien. Il tombe pas sous le raisonnement discursif ordinaire, si bien que la *noësis* ne peut être que de nature intuitive. Il s'agit à proprement parler de cette partie de la raison qui, lorsqu'elle s'émancipe des sens, peut être conduite à la révélation et à la contemplation de l'Idée du Bien, du Beau et du Vrai, qui participent d'une seule et même essence. C'est pourquoi nous la traduirons par raison intuitive ou encore, contemplative. Socrate dit de celui qui en use « [qu'] il a recours à la construction d'hypothèses sans les considérer comme des principes, mais pour ce qu'elles sont, des hypothèses, c'est-à-dire des points d'appui et des tremplins pour s'élancer jusqu'à ce qui est anhypothétique, jusqu'au principe du tout²⁸ » Comme nous allons le voir dans le chapitre suivant, cette idée est assez proche de ce que Kant nomme la Raison pure, mais en un sens plus contemplatif, quasi mystique à certains égards. La philosophie moderne a dépouillé ce qu'il y avait de poétique dans la pensée de Platon, sans toutefois perdre la notion « d'intuition » ou « d'évidence ».

En conclusion, je dirais que le mouvement essentiel opéré par Platon, sur la question qui nous préoccupe, a consisté à rompre l'identité, admise par Protagoras, entre la vérité et le verdict des organes sensoriels. Dès lors que le langage n'est plus une simple sensation, il doit s'employer, avec la raison, à trouver l'être des choses dans leur dimension purement intelligible. Cette recherche passe par les impressions, et porte sur elles, mais se réalise dans le jugement, dont la destination est de dire leur unité, leur universalité, ou en d'autres termes, leur être même. Dans la vision la plus élevée du monde, qui est celle de la contemplation intellectuelle, il n'y a pas de distinction entre les idées du Beau, du Bien, du Vrai. Toutes les formes de la nature se confondent. Elles se répondent les unes les autres, se complètent, dans une unité parfaite. C'est la raison pour laquelle la plus haute forme d'expérience esthétique, chez Platon, est aussi une expérience morale et scientifique. Il y a derrière le monde quelque chose d'universel qui efface les frontières perçues par les sens entre les objets. Lorsque le sujet tend vers cette unité, il se transforme de l'intérieur. Il se purifie. Il devient un être humain à part entière.

28. *République*, 511b.

Dès les balbutiements de la philosophie, Platon distingue trois espèces de jugement :

1. Le *jugement d'expérience* est synthétique. Il s'exerce par la faculté de la sensation et consiste à observer le monde sensible.
2. Le *jugement rationnel* est analytique. Il s'exerce par la faculté de la raison et consiste à trouver l'essence des choses — leurs définitions pourrait-on dire.
3. Le *jugement vrai sans justification*, quant à lui, permet de réunir sous un concept différentes réalités sensibles, mais sans en détenir un critère explicite. Contrairement aux autres formes de jugement, Platon ne lui associe pas de faculté particulière. Il se contente de dire qu'il s'agit de l'exercice même de l'âme humaine.

C'est sur ce jugement sans justification, comme on va le voir dans le prochain chapitre, que Kant développe sa théorie de l'Entendement humain.

Chapitre 2

Le Jugement chez Kant : La Genèse de l'Entendement

J'ai montré dans la section précédente que Platon identifiait quatre types d'aptitudes : l'eikasia (imagination), la pistis (croyance), la diánoia (discursivité) et la noësis (intellection). Les deux premières ressortent de la faculté de la Sensation tandis que les deux autres relèvent de la Raison. Dans le monde sensible, les jugements sont établis au moyen des sens, de l'expérience, de la mesure. Dans l'intelligible, en revanche, la vérité est une question de déduction rationnelle, qui part des premiers principes et qui s'élève, de proche en proche, vers des vérités supérieures.

Pour illustrer les différences existant entre le jugement sensible et le jugement rationnel, Platon donne l'exemple du géomètre qui trace un triangle sur le sol et qui mesure que la somme des angles fait 180° . Tant qu'il se base sur le sens de la vue, il reste dans le domaine du sensible. Il observe des récurrences, il cherche des contre-exemples, il établit des corrélations. Son jugement peut être vrai, mais ce ne sera que par accident. La manière idoine de poser un jugement, d'après Platon, est de fournir une preuve basée sur les définitions. C'est en partant du concept de triangle et du concept d'angle, en regardant ce qui est contenu dans leur essence, que l'on arrive à conclure, de façon certaine, que le triangle respecte telle ou telle autre propriété géométrique.

On peut dire, sans prendre trop de risque, que les Éléments d'Euclide constituent sans doute le modèle de la vérité chez Platon. Les Éléments illustrent en effet l'idée qu'un triangle, qu'un angle, qu'un cercle, sont d'une autre nature que les formes géométriques tracées dans le sable ou sur le papier granuleux. Ils ont une essence, une définition, de laquelle découlent des propriétés nécessaires, démontrables et éternelles.

Il s'ensuit que deux espèces de jugements peuvent être dégagés chez Platon : ceux qui sont vrais sans justification logique, parce que fondés sur l'expérience du monde ; et ceux qui sont vrais et munis de l'explication vraie.

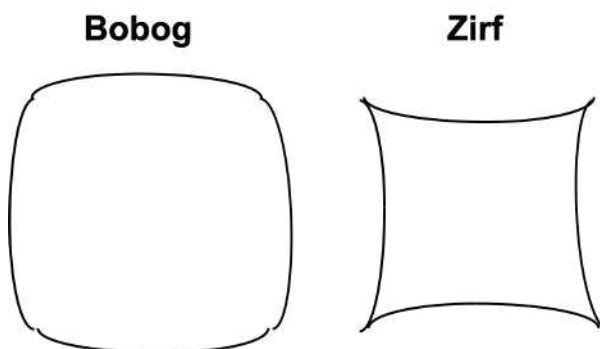


FIGURE 2.1 – Deux figures géométriques inconnues.

Ces deux espèces de jugement préfigurent la distinction existant entre les jugements *synthétiques*, basés sur l'expérience du monde (exemple : « est-ce que cet objet ressemble à celui-ci ? »), et les jugements *analytiques*, basés sur la raison seule (exemple : les démonstrations mathématiques).

Il existe cependant une voie du milieu, qui est présente dans le texte de Platon, quoi qu'elle ne soit pas objectivement thématifiée et distinguée des deux autres. Cette voie concerne le jugement vrai sans justification ; un jugement qui, en restant de l'ordre du concept rationnel, se rapporte néanmoins à des objets sensibles. Ce jugement intervient lorsque le sujet pensant cherche à mettre en relation, non pas deux idées l'une avec l'autre (l'idée du triangle et l'idée d'un angle), ni deux objets sensibles (le dessin d'un triangle et celui d'un angle) ; mais une idée avec une forme sensible. L'on peut penser par exemple à quelqu'un qui verrait un dessin dans le sable et qui dirait, avec une emphase digne d'Archimède découvrant les lois de l'hydraulique : « regarde, c'est un triangle ! »

Ce jugement dont je parle n'est pas équivalent (1) au jugement rationnel ou analytique ; il n'est pas non plus équivalent (2) à un jugement synthétique basé purement sur les sens. Pour s'en rendre compte, j'aimerais discuter une petite expérience de pensée de mon invention.

2.1 L'Abstraction intuitive

Supposons que l'on vous présente deux figures géométriques inconnues ayant pour nom le « Bobog » et le « Zirf » (voir figure 2.1). Sur la base de ce seul exemple, on vous demande de classer une pile de dessins selon qu'il s'agit de Bobog ou de Zirf, sans vous donner aucune définition précise. Pour vous aider, votre collègue a pris le soin d'accompagner chaque figure d'une annotation disant soit « Bobog », soit « Zirf », mais vous ne savez pas si cette annotation est correcte ou non (voir figure 2.2).

La première stratégie consiste à parcourir chaque image une à une en

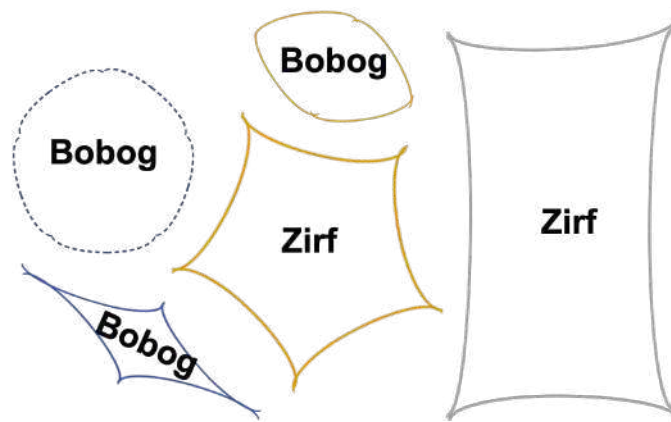


FIGURE 2.2 – Un ensemble de figures géométriques qui ont été annotées par un collègue. S'agit-il de Bobogs ou de Zirfs ? L'annotation de votre collègue n'est pas forcément exacte.

prêtant attention aux annotations. S'il s'agit du mot « Zirf », vous mettez le dessin dans la pile de gauche. S'il s'agit du mot « Bobog », vous la mettez dans la pile de droite. Cette façon de faire est uniquement rationnelle. Vous n'avez pas besoin, pour cette méthode, d'identifier réellement les formes qui sont dessinées. La résolution, du reste, peut aisément être confiée à un algorithme. Des limitations évidentes se rencontreront toutefois si l'annotation est fausse, ou si un dessin est accompagné d'un mot non répertorié, comme « Bozif » ou « Zirob ».

La deuxième stratégie consiste à mettre tous les dessins à plat devant vous, de sorte à les voir en leur entier, et à regrouper ceux qui présentent, entre eux, le plus de similarités statistiques. Vous faites donc deux piles en vous fiant à votre intuition. Si vous êtes un être humain normalement constitué, vous observerez que certaines formes présentent des contours arrondis, tandis que d'autres formes ont des arrêtes accidentées. Vous les mettez donc dans des piles séparées. Il ne vous restera plus qu'à déterminer alors quelle est la pile qui correspond au Bobog qu'on vous a donné en exemple et quelle est celle qui correspond au Zirf. Pour résoudre cette dernière étape, vous vous référerez, là aussi, aux ressemblances statistiques. L'ensemble de la méthodologie consiste donc à explorer intelligemment les deux questions que voici : Quelles sont les formes présentant le plus de ressemblances entre elles ? Quelles sont les formes présentant le plus de ressemblances avec le Bobog et le Zirf qu'on vous a présenté en exemple ?

Cette approche que je viens de décrire n'utilise pas les annotations de votre collègue. Il peut cependant arriver qu'elle conduise à des résultats inattendus. En regardant quelles formes présentent, entre elles, le plus de

ressemblances statistiques, vous pourriez être amené à regrouper celles qui sont réalisées à l'encre bleue et celles qui sont faites à l'encre jaune. Il se pourrait aussi que vous mettiez ensemble les formes de grande taille avec celles de petite taille, ou que vous sépariez celles faites avec des traits pleins de celles faites avec des pointillés. Il n'existe à vrai dire pas une seule et unique réponse valide. Autrement dit, l'association statistique n'est pas du même ordre que l'approche rationnelle vue précédemment. Au lieu de reposer sur des jugements rationnels, définis à l'aide de règles linguistiques univoques, le jugement synthétique mobilise des considérations basées sur l'expérience. Les réseaux de neurones procèdent de cette manière et, faute d'un entraînement intensif, il leur arrive de se tromper.

La troisième stratégie, pour finir, consiste à regarder les dessins l'un après l'autre et à se demander, dans chaque cas, s'il s'apparente à l'idée générique du « Bobog » ou à l'idée générique du « Zifr ». Cette idée générique est obtenue en retranchant du modèle de départ tout ce qui ne semble pas essentiel. Vous faites donc abstraction de la taille, de la couleur et de l'épaisseur des traits, pour ne vous intéresser qu'à ce qui semble constituer le plus important contraste : les Bobogs sont constituées de courbes bombées vers l'extérieur (courbes convexes), tandis que les Zirfs sont construits à l'aide de courbes bombées vers l'intérieur (courbes concaves). Cette troisième stratégie vous invite donc à construire dans votre tête un archétype généralisé du Bobog et du Zifr en abstrayant les propriétés significatives. Cette démarche repose sur des jugements qui sont certes vrais, mais pas totalement justifiés. Il y a certes une volonté de dépasser les simples constats de ressemblances expérimentales, mais sans pouvoir s'appuyer sur la définition rationnelle de ce qu'est un Bobog ou un Zifr.

Je crois déjà entendre le lecteur formuler une objection capitale : la troisième stratégie diffère-t-elle vraiment de la deuxième ? Cette capacité d'abstraction que j'ai décrite, et que j'ai distinguée de ce que font les réseaux de neurones, ne serait-elle pas le produit d'associations statistiques ? Si l'on entraînait une machine à compter le nombre de côtés, tout comme l'enfant apprend à reconnaître des formes géométriques en jouant avec des cubes en bois, n'aboutirait-on pas à quelque chose de tout à fait analogue ? Existe-t-il réellement dans l'abstraction géométrique un élément qui ne relève pas du jeu de l'expérience statistique ?

Cette objection, qui est d'une importance névralgique, recevra diverses réponses tout au long de ce livre (voir notamment le chapitre [8.1](#)). On peut notamment insister sur la différence d'apprentissage. Pour créer un algorithme capable d'identifier des objets complexes, ou de produire du texte comme ChatGPT-4, il faut l'entraîner avec plus de données qu'un être humain n'en verra jamais au cours de sa vie. Les résultats, bien qu'impressionnants, ne rivalisent pas toujours avec le jugement humain. Un même célèbre, qui a défrayé les réseaux sociaux en 2022, montrait comment un algorithme peut confondre un muffin au chocolat avec le visage larmoyant d'un chi-



FIGURE 2.3 – Les réseaux de neurones, faute d'un entraînement intensif, commettent des erreurs d'identification plus fréquentes que l'être humain adulte face à ce genre d'images.

huahua... ou un vulgaire trognon de pomme avec les yeux sagaces d'une chouette effraie (voir figure 2.3). Il semblerait que les réseaux de neurones, bien qu'ils surpassent les performances humaines à toutes sortes de tâches, ne reposent pas nécessairement sur les mêmes principes fondamentaux que la cognition humaine. Celle-ci, en effet, bien que faillible dans ses analyses logiques, est capable de comprendre des concepts sur base d'un tout petit nombre d'échantillons.

2.1.1 Le Jugement synthétique *a priori*

Les catégories manipulées par l'esprit humain sont des *concepts intuitifs*. Le mot *concept* indique qu'il y a bien quelque chose, là-dedans, qui est de l'ordre de la raison, mais le mot *intuitif* nous rappelle qu'il y a également présence d'une inspection à l'aide de nos sens. Le type de jugement permettant d'isoler des concepts intuitifs, quoiqu'il soit assez banal dans la vie quotidienne, présente quelque chose de tout à fait extraordinaire pour la philosophie. C'est un jugement qui se rapporte non pas à des objets de l'expérience concrète (des formes géométriques observables), mais à des objets d'expérience possible (l'idée générique du Zirf et du Bobog).

Qu'est-ce que je veux dire par là ? Que ce type de jugement repose sur un important *a priori* : il suppose que les Bobogs et les Zirfs que nous avons sous les yeux ne sont que des exemplaires possibles d'une structure plus générale. Il est donc supposé que ce que nous déterminons de vrai et de faux à propos de ces cas particuliers nous apprend quelque chose de tous les objets

analogues. Pour le dire encore autrement, nous sommes incapables d'observer une forme géométrique aussi aléatoire qu'un Bobog sans nous rapporter, en imagination, à toutes sortes d'autres formes géométriques possibles. Nous comparons le Bobog aux cercles, aux carrés, aux triangles de notre mémoire et nous cherchons à déterminer des ressemblances, des contrastes, des points de rupture.

Ce phénomène dont je parle est rendu particulièrement visible dans la vie sociale pratique. Lorsque vous tentez de cerner le caractère d'une personne que vous venez de rencontrer, vous vous adonnez à toutes sortes de supputations. Pour rendre mon propos tangible, supposons que vous ayez matché avec un jeune homme ou une jeune femme sur votre application de rencontre favorite et que vous lui ayez donné un premier rendez-vous. Votre esprit, dès cet instant, est accaparé par la question de savoir qui est cette personne. Correspond-elle à l'image que vous vous en êtes faite en lisant son profil ? Lorsque vous la voyez pour la première fois, que vous l'entendez parler, que vous découvrez sa logique interne, vous la comparez aux individus que vous connaissez déjà. Il vous serait peut-être même impossible de cerner la personnalité d'un inconnu sans l'assimiler à toutes sortes de personnes qu'elle n'est pas, mais qui pourtant lui ressemblent. Vous la comparez à des personnes réelles que vous avez connues, mais également à des personnages de fiction, à des archétypes. Ces derniers sont si présents dans notre imaginaire que la langue est truffée d'antonomases : on dit d'un avare que c'est un Harpagon, d'un séducteur que c'est un Don Juan ou un Casanova, d'un génie qu'il est un Da Vinci ou un Einstein, etc. Une personne lascive, pour ceux qui connaissent la littérature russe, est comparée à Anna Karénine. C'est un équivalent de l'Emma Bovary des français, avec une touche « poétesse maudite » bien russe.

C'est pourquoi on pourrait dire du jugement humain, lorsqu'il manipule des concepts intuitifs, qu'il est certes *synthétique*, mais *a priori*. Il est *synthétique* car il étend notre connaissance en faisant des associations nouvelles (il nous éclaire sur la personnalité de quelqu'un, sur les propriétés de certaines formes géométriques, etc.). Il est toutefois *a priori* car il dépend d'un ensemble de préjugés : nous supposons d'avance que la personne que nous observons ou que la figure géométrique que nous étudions participe d'un archétype. Cet archétype dépend de notre façon de structurer l'information et de nous représenter le monde. Il est donc dépendant de nos concepts culturels et de nos prédispositions neurologiques.

Le jugement synthétique *a priori* nous éclaire parfois davantage sur nos filtres perspectifs, et sur notre façon de lier l'information, que sur les objets auxquels il se rapporte. C'est pourquoi, lorsqu'une personne A dit quelque chose à propos d'une personne B (que B est avare, séductrice ou manipulatrice...) on en apprend tout autant sur l'univers mental de cette personne A que sur les comportements de B. Pour le dire avec une pointe de poésie, le jugement synthétique *a priori* est cet espace dans lequel des associations

sont faites entre des choses connues et inconnues, entre des objets réels et des objets possibles, entre des vérités expérimentales et des *a priori*. Il rend visible la liaison existant entre nos concepts. Il révèle la structure profondément interdépendante des idées que nous manipulons.

L'exemple historique le plus simple et le plus élégant de jugement synthétique *a priori* peut être trouvé dans la démonstration de Thalès, comme quoi « le diamètre d'un disque coupe ce disque en deux parts égales ». À première vue, cette proposition n'a rien d'extraordinaire. S'échiner à prouver un tel résultat s'apparente à vouloir enfoncer une porte ouverte. Ce qui est pourtant extraordinaire avec cette preuve, ce n'est pas tant la conclusion que le procédé. Pour la première fois de l'histoire des mathématiques, l'on observe une volonté d'établir une démonstration valide pour tous les disques et tous les diamètres *possibles*. Loin donc de mesurer que la propriété est exacte en dessinant des cercles avec un compas, on cherche à universaliser le raisonnement de telle sorte à ce que les objets que l'on manipule ne soient plus ce cercle-ci ou ce cercle-là, mais le cercle en général — le cercle *archétypal*.

La preuve suggérée par Thalès recourt à ce qu'on appelle une preuve par l'absurde. Elle consiste donc à supposer que ce que l'on cherche à démontrer est faux, puis à conclure que cela conduit à une contradiction. En vertu du principe d'après lequel soit une chose est vraie, soit une chose est fausse — il s'agit là du principe du tiers exclu — on est alors forcé d'admettre que la chose à démontrer, ne pouvant être fausse, doit être vraie. On trouve une bonne explication de la preuve de Thalès dans le podcast de Intellectual Mathematics [\[1\]](https://intellectualmathematics.com/blog/first-proofs-thales-and-the-beginnings-of-geometry/).

Supposons que le diamètre ne divise pas le cercle en deux moitiés égales. Très bien, nous avons donc une ligne qui passe par le centre d'un cercle, et ce cercle est coupé en deux parties. Et nous supposons que ces deux parties ne sont pas identiques. Prenons une des parties et retournons-la sur l'autre, comme si vous pliez une omelette ou une crêpe. Nous avons supposé que les deux parties n'étaient pas égales, donc en superposant l'une sur l'autre, elles ne correspondent pas parfaitement. Il doit donc y avoir un endroit où l'une des deux parties dépasse l'autre.

Maintenant, tracez un rayon dans cette direction, depuis le centre du cercle jusqu'au point du périmètre où les deux moitiés ne correspondent pas. Ainsi, un rayon est plus long que l'autre. Mais cela signifie que la figure n'était pas un cercle au départ. Un cercle est une figure où tous les points du périmètre sont à égale distance du centre dans toutes les directions. C'est justement

1. Traduit depuis l'anglais depuis *First proofs : Thales and the beginnings of geometry*, May 16, 2020, <https://intellectualmathematics.com/blog/first-proofs-thales-and-the-beginnings-of-geometry/>

cela, être un cercle.

Dans cette preuve, l'être humain ne cherche pas simplement à établir des relations statistiques entre des objets dessinés sur le papier. Plutôt que de regarder les cercles et les diagonales avec ses yeux, il recourt à des images génériques. L'image générique d'un disque que l'on replie sur lui-même ne désigne pas une expérience *particulière*, mais une expérience *possible*. Ce que je veux dire par là, c'est que l'expérience de pensée au travers de laquelle vous effectuez cette opération est valide pour tous les cercles existants ou imaginables. Vous savez donc que, même si le cercle que vous vous représentez a une certaine dimension, cette dimension est sans importance. Autrement dit, vous avez une notion universelle du cercle qui vaut pour tous les cercles représentables.

Pour cette raison, on peut dire que les conclusions auxquelles vous parvenez sont des *a priori* dans le double sens du terme : universel et arbitraire. Universel, tout d'abord, puisqu'elles sont valables dans toutes les expériences possibles. Quel que soit le diamètre du cercle, sa couleur, son épaisseur, sa position sur la page, vous obtiendrez invariablement le même résultat.

Ces conclusions sont également arbitraires, car on n'est pas sûr qu'un cercle existe dans la réalité. Si vous dessinez un cercle sur le papier, vous observerez qu'il n'est pas tout à fait circulaire, qu'il existe des bosses, des dérapages, des inexactitudes, si bien que la démonstration de Thalès ne fonctionnera pas dans la pratique. Il est donc quelque peu arbitraire de penser que ce que vous avez dessiné soit un cercle, et non un nuage de points. On peut estimer que d'autres cultures, bien différentes de la nôtre, rapporteraient les observations à d'autres idées génériques, à d'autres figures abstraites, que celle du cercle que nous connaissons. Leur géométrie ne serait pas la géométrie euclidienne à laquelle nous sommes familiers, mais un ensemble de filtres perceptifs aussi étrange qu'inimaginable. Je reviendrai plus tard sur cette idée (voir notamment le chapitre 4.3).

Pour résumer ce qui vient d'être dit à propos des jugements synthétiques *a priori*, la suite de cette thèse consistera à mettre à l'épreuve (avec des résultats mitigés) l'idée suivante :

Thèse 2. *Il existe chez l'être humain une faculté de jugement qui n'est imitable ni par le truchement de l'expérience pure (associations statistiques) ni par le truchement de la logique pure (manipulation de règles formelles), mais qui consiste en une forme d'abstraction permettant de subsumer les objets de l'expérience sensible sous une catégorie rationnelle. Dans la théorie kantienne (que je critiquerai ultérieurement), ce jugement est appelé le jugement synthétique a priori. Il ne se rapporte ni à des objets réels, ni à des définitions abstraites, mais à des objets possibles.*

2.1.2 Les Limites du platonisme

La pensée de Platon, comme on l'a vu, est influencée par la méthode de démonstration géométrique. Elle met à l'honneur les démonstrations *a priori* de Thalès et d'Euclide. Elle ne distingue cependant pas de troisième niveau entre les jugements purement synthétiques et les jugements purement analytiques. Si le père de l'Académie d'Athènes ne donne pas de nom spécifique aux jugements synthétiques *a priori*, c'est parce qu'il estime que la réalité du monde est peuplée d'essences. Pour lui, seule est vraie l'Idée du triangle, et non les ersatz qui se dessinent dans le sable. Les Idées, bien qu'elles soient représentées par les sens, ne tombent pas pour autant sous l'emprise des sens. Elles ne sont accessibles que par le truchement de notre intellect. C'est pourquoi, dans le platonisme, toutes les choses du monde, aussi bien les formes dans le sable que le bruit des vagues en marée haute, le vent hurlant, les embruns, le sel et le vacarme du ressac ; toutes, en dernier ressort, ne font rien que participer de l'expression, dans le monde, des essences universelles. Le monde sensible est la manifestation concrète, quoique déformée, d'une réalité idéale.

Pour exprimer autrement la pensée de Platon, il doit exister dans le monde intelligible (topos intelectos) une Idée du triangle, une Idée des vagues, de la marée, du vent hurlant, des embruns, du sel et du ressac. Ces Idées ne sont toutefois pas comparables aux objets du monde sensible qui nous tombent sous les yeux. Ce ne sont pas non plus des objets possibles comme dans la théorie de Kant. Les idées de Platon sont des essences pures, des formes idéales — des définitions exactes. Notre âme a pu contempler ces essences avant de naître dans un corps, et aussi en garde-t-elle le souvenir. Lorsque cette âme dit d'une forme confusément inscrite dans le sable « ceci est un triangle », elle ne fait rien d'autre que retrouver un souvenir primordial, un songe perdu, une mémoire ancienne et inaltérable.

La méthode d'apprentissage qui permet à l'âme d'accoucher des vérités qui sont en elle est appelée la maïeutique. Elle consiste à poser les bonnes questions de façon à ce que l'être humain, par l'usage de sa seule raison, se hisse au plus haut degré d'affûtage des définitions, des concepts analytiques et des relations les unissant. Platon en donne un exemple dans le dialogue du Ménon, où Socrate, en interrogeant adroitement un esclave, parvient à guider ce dernier vers la compréhension de vérités géométriques.

C'est pourquoi le genre de jugement qui consiste à dire « regarde, c'est un triangle ! » n'intéresse pas Platon au premier chef. Il tombe du côté de la Pistis (la croyance) et n'a pas de statut épistémologique privilégié. Ce jugement est pourtant discuté dans le Théétète ; je l'ai indiqué à plusieurs occasions. C'est même ce jugement qui est à l'origine de la notion « d'Entendement ». Ce mot, qui apparaît sous la plume de Descartes, Spinoza, Hume ou Locke, désigne une faculté occupant une place intermédiaire entre la Sensibilité, concentrée sur les objets sensibles, et la Raison — qui contemple les

pures formes intelligibles.

Cette notion « d'Entendement » nous interpelle au premier plan, car il semblerait que ce soit précisément un point sur lequel l'intelligence artificielle présente encore quelques pénibilités. S'il est facile d'entraîner une machine à faire des déductions logiques, en tirant d'une première idée rationnelle une conclusion toute aussi rationnelle ; s'il est également permis de demander à l'algorithme de repérer des patterns au sein des données sensibles ; il reste cependant difficile de généraliser des observations sensibles de façon à les manipuler de façon abstraite. Autrement dit, l'intelligence artificielle, en comparaison avec l'être humain, a du mal à former des catégories manipulables par la pensée rationnelle.

Dans la section de l'introduction intitulée *Le QI artificiel*, j'ai illustré certains types de problèmes pour lesquels les machines restent médiocres. Ces problèmes, inventés par François Chollet, s'apparentent à des problèmes de Raven. Ils consistent à généraliser certains patterns à partir d'un petit ensemble d'exemples (voir figure 2.4). Les tâches qui exigent des structures logiques flexibles, et ne reposent pas purement et simplement sur l'imitation statistique de vastes ensembles de données, sont susceptibles de faire obstacle aux paradigmes de programmation actuels. Ces structures logiques que j'ai en vue consistent souvent à distinguer la cause de l'effet, à distinguer l'unité de la pluralité, à réaliser des symétries, etc. On pourrait dire que ces opérations de pensée sont des espèces de jugements, qui consistent à décomposer l'image initiale, prise comme un tout, en éléments individuels significatifs.

En raison de son haut degré d'abstraction, ce genre de jugement semble occuper un milieu entre ce que Platon appelait la Pistis (croyance) et la Dianoia (logique déductive). C'est un jugement qui est certes synthétique, au sens qui regroupe des ensembles de pixels sous une même unité, mais pas totalement *a posteriori* (qui dépendrait donc uniquement de l'expérience, sans faire intervenir la raison). Loin en effet de se fonder uniquement sur l'expérience, il doit s'appuyer en partie sur des lois abstraites, des propriétés de symétries, des intuitions spatiales et temporelles — qui sont à mille lieues des vastes ensembles de corrélations statistiques des big data.

Dans la mesure où ce n'est ni un jugement synthétique *a posteriori* (il ne repose pas uniquement sur l'expérience), ni un jugement purement analytique (il ne repose pas uniquement sur des définitions), il semblerait qu'il dépende d'un autre ingrédient que ceux que nous avons vus en étudiant la pensée de Platon. Dans les pages qui vont suivre, j'aimerais suggérer que cet ingrédient pourrait être comparable au jugement synthétique *a priori*, introduit par Emmanuel Kant dans sa monumentale Critique de la Raison pure.

Ce choix de vous parler de Kant ne saurait être arbitraire. Non seulement il s'agit du philosophe qui a le plus insisté sur le rapport du jugement humain à l'expérience, en mettant en lumière l'existence d'une capacité d'abstraction

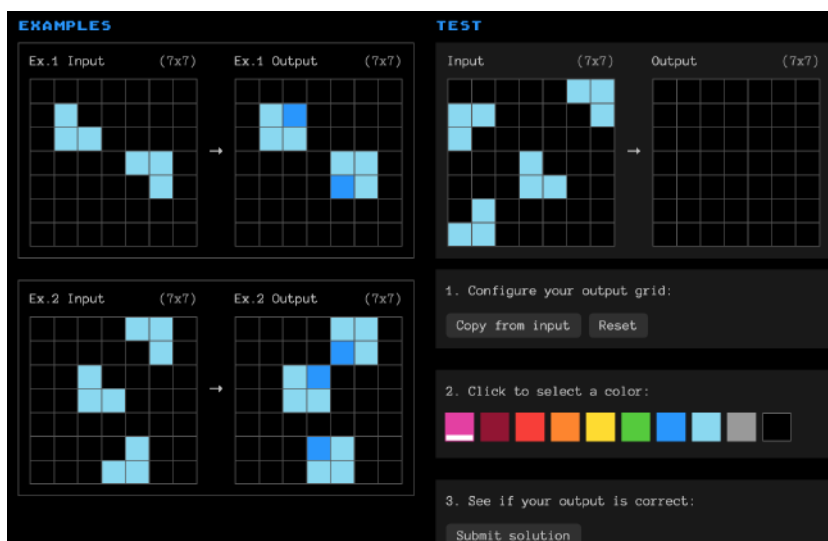


FIGURE 2.4 – Dans cette matrice de Raven, on donne à la machine des morceaux de carrés dont il manque un pixel et on lui montre, au moyen de deux seuls exemples, qu'elle doit compléter le carré avec un pixel bleu. La machine doit ensuite appliquer ce qu'elle vient d'apprendre à de nouvelles situations.

intuitive, mais il est en outre connu pour avoir établi un étroit dialogue avec la théorie de Platon. Il a tenté de sortir la philosophie de cette espèce de rationalisme spéculatif, de ces limbes pétrées d'idéalité où les dieux grecs semblaient l'avoir engendrée.

Dans les prochains chapitres, je vais me concentrer tout particulièrement sur la Critique de la Raison pure, m'efforçant de relever plusieurs passages se référant au Théétète. Je tâcherai d'introduire la définition du jugement synthétique *a priori* dans une analyse comparée avec les dialogues socratiques, après quoi je montrerai comment ceux-ci participent à l'introduction d'une nouvelle faculté humaine : l'Entendement. Ces discussions prépareront le terrain, je l'espère, à une discussion plus approfondie sur mon hypothèse principale : le jugement synthétique *a priori* est-il précisément le hiatus, l'inconnue, la carence qui manque à l'intelligence artificielle pour résoudre des exercices d'abstraction similaires aux tests ARC de François Chollet ?

Il est toutefois important de noter que ma thèse ne s'inscrira pas, purement et simplement, dans la théorie de Kant. J'introduis ici le concept de jugement synthétique *a priori* de manière à construire, ultérieurement, une théorie de l'esprit qui soit capable de rendre compte du pourquoi nous pouvons nous faire une image mentale du « cercle en général » ou du « diamètre en général », et comment nous pouvons manipuler ces concepts pour établir des relations.

2.2 Kant vs Platon : Le Procès des Idées

Kant reprend la philosophie à l'endroit précis où Platon l'avait trouvée, mais plutôt que d'étendre aveuglément la dialectique socratique aux questions d'ordre métaphysique, plutôt que de répondre à la question de savoir ce qui rend possible une science (universelle) du sensible et de ses transformations (comme dans le *Théétète*), il souhaite adresser préalablement un soupçon d'ordre méthodologique : qu'est-il possible de connaître à l'aide de la raison ? Il serait oublieux, à ses yeux, de conquérir le monde intelligible sans comprendre préalablement ce qui nous en donne l'accès. Car avant de faire usage d'un outil, fût-il le plus puissant dont on dispose, il convient d'en produire l'analyse. C'est sans doute ce qui lui fait dire que :

« S'il était venu à l'esprit de quelque ancien de poser seulement cette question, elle aurait opposé à elle seule une puissante barrière à tous les systèmes de la raison pure qui se sont élevés jusqu'à nos jours, et elle aurait épargné bien des tentatives inutiles, auxquelles on s'est livré aveuglément sans savoir proprement de quoi il s'agissait². »

Quelles sont exactement ces tentatives que Kant condamne ? Quels sont les griefs qu'il exprime à l'endroit de la philosophie occidentale et de la métaphysique en particulier ? Pour s'en faire une idée juste, peut-être n'est-il pas inutile de rappeler que Platon a introduit la Raison de façon à permettre un nouveau type de jugement, qui soit à la fois universel et synthétique. Ce faisant, il a conduit la philosophie dans le royaume de la pensée pure — une pensée qui ne prend plus pour objet des impressions sensibles, mais des concepts universels comme celui des nombres. Cela lui a permis d'envisager une connaissance de l'intelligible qui, à l'instar des mathématiques, s'exprime au travers d'images sensibles, mais sans dépendre d'elles. Une connaissance qui dépasse le sensible et s'élève progressivement à l'anhypothétique.

La conception de Platon se fait hélas au prix d'un certain dogmatisme : l'existence des Formes et la transcendance de l'Idée du Bien (dont on peut se demander à quoi elle ressemble exactement). S'il faut donc être redevable à Platon d'avoir insisté sur l'existence d'une connaissance apriorique, il faut également lui reprocher de n'avoir pas fourni une réelle *pensée* de la pensée pure. Selon quel critère peut-on déterminer si une connaissance est à l'image d'une Idée ou si elle n'est qu'une opinion dévoyée ? Quelles sont les propriétés des pensées pures et comment s'articulent-elles les unes aux autres ? Quelles sont, en d'autres termes, les règles du raisonnement ? Platon s'est certes interrogé sur l'existence de l'opinion fausse, sur ce qui la distingue de l'opinion vraie, mais de façon chaque fois assez aporétique. Il s'est comme empressé de construire un édifice rationnel à même d'élever la

2. KANT, *Critique de la Raison pure*, trad. de BARNI, Introduction V, note (a) de la page 58.

connaissance humaine au faîte de sa puissance, sans chercher préalablement à en vérifier les fondations.

Dans son introduction à la *Critique de la raison pure*, Kant redouble d'ironie à l'égard de la théorie des Idées : « La colombe légère, qui, dans son libre vol, fend l'air dont elle sent la résistance, pourrait s'imaginer qu'elle volerait bien mieux encore dans le vide. C'est ainsi que Platon, quittant le monde sensible, qui renferme l'intelligence dans de si étroites limites, se hasarda, sur les ailes des idées, dans les espaces vides de l'entendement pur^[3] » Bien qu'il n'explicite pas le sens de ces quelques lignes, celui-ci se donne à voir dans un nombre important d'autres références plus ou moins directes à la théorie des Idées^[4].

Il est d'ailleurs vraisemblable que la « colombe légère » renvoie directement à l'analogie du colombier du *Théétète*^[5]. Nous l'avons déjà exposée dans la section intitulée *Le Jugement Vrai* (section ??), mais sans en aborder tous les détails. Comme elle n'est pas de moindre importance, je vais prendre le soin d'en discuter ici plus précisément les limitations, en adoptant tout au long de ma démarche un point de vue kantien. Ce sera l'occasion de mettre en évidence les différences fondamentales qui séparent, selon moi, les deux auteurs. Cela permettra aussi d'introduire le projet de la *Critique de la Raison pure*, en ce qu'il a de novateur concernant le point de vue posé sur les propositions synthétiques.

2.2.1 La Métaphore du colombier

Dans le *Théétète*, une théorie de la connaissance est exposée au moyen de la métaphore du colombier. Cette allégorie a pour vocation de rendre compte de la possibilité de commettre des erreurs de jugement dans la pensée pure. Pour rappel, elle consiste à assimiler les savoirs et les ignorances à des oiseaux sauvages^[6]. Apprendre dans ce contexte, c'est capturer l'oiseau approprié et le mettre dans la cage étiquetée d'un grand colombier. Se rappeler et

3. *Ibidem*, Introduction III, pp. 52-53.

4. Voir notamment la première section de la dialectique transcendantale : « Platon se servait du mot idée de telle sorte qu'on voit bien qu'il a entendu par là quelque chose qui non seulement ne dérive pas des sens, mais dépasse même les concepts de l'entendement dont s'est occupé Aristote, puisqu'on ne saurait rien trouver dans l'expérience qui y corresponde. Les idées sont chez lui les types des choses elles-mêmes, et non de simples clefs pour des expériences possibles comme les catégories. Dans son opinion, elles dérivent de la raison suprême, d'où elles ont passé dans la raison humaine, mais celle-ci ne se trouve plus dans son état primitif et ce n'est qu'avec peine qu'elle peut rappeler aujourd'hui, bien obscurcies, ses anciennes idées, par la réminiscence (qui s'appelle la Philosophie) » (KANT ^[101], *op. cit.*, p. 311, voir aussi pp. 312-315).

5. Voir *Théétète* 197b-200e.

6. La théorie du colombier est élaborée tout au long des dialogues. Dans une première version (*Théétète* 197b-197e), chaque colombe porte un savoir, et l'erreur consiste donc à confondre un oiseau avec un autre. Socrate fait alors remarquer quelque chose de gênant : l'erreur vient toujours de ce que le sujet sait quelque chose, mais qu'il confond ce qu'il connaît, et non pas de l'ignorance pure.

appliquer un savoir qui a été appris, c'est dès lors rechercher dans sa cage l'oiseau détenteur de ce savoir.

Selon cette conception, l'erreur survient lorsque l'on prend un oiseau pour un autre, ou que l'on prend une ignorance à la place d'un savoir. Cette théorie suppose donc qu'il existe deux genres de raisonnements : les premiers sont des jugements se rapportant à des impressions sensibles, les seconds sont des jugements de la pensée pure. Aux dires de Socrate, et à travers lui de Platon qui tient la plume, une opération mentale telle que « $5 + 7 = 12$ » est un acte de la pensée pure⁷, qui en tant que tel ne se rapporte à aucune intuition sensible.

Il peut être utile de se demander où nous entraîne une telle théorie des pensées pures ? Dans la mesure où certaines colombes, au lieu de porter des savoirs authentiques, transportent des ignorances, il doit exister quelque part un savoir se rapportant à ces colombes, et permettant de dire si ce qu'elles détiennent est vrai ou faux. Cela ne signifie pas autre chose que ceci : pour chaque colombe, il doit être possible de savoir si ce qu'elle dit est valide en se référant à l'information d'une autre colombe. La pertinence de ce que nous a dit cette deuxième colombe, à son tour, doit faire l'objet d'une vérification à l'aide d'une troisième colombe, et ainsi de suite. À ce titre, l'on est bien vite entraîné vers une régression infinie.

En des termes plus contemporains, le problème soulevé par Socrate est donc celui de savoir ce qui rend possible la réflexivité de la connaissance. Soit l'on admet que toute pensée, par cela même qu'elle est une pensée de quelque chose, est nécessairement vraie, un peu à la façon de Parménide, mais dans ce cas l'on est mis en mal d'expliquer l'origine de l'erreur — et l'on risque de sombrer avec Protagoras dans les méandres du relativisme. Soit au contraire l'on admet qu'il puisse exister des pensées fausses, mais alors il doit aussi y avoir une pensée des pensées fausses, puis une pensée des pensées des pensées fausses, et ainsi de suite, de sorte que la vérification finale de la validité ne puisse jamais être fournie. Rechercher la vérité dans ces conditions, c'est poursuivre le terme ultime d'une suite de conditions qui s'enchaînent sans fin. C'est vouloir remonter à l'inconditionnel ou, dans les termes de Platon, à l'anhypothétique.

Cette objection est soulevée par Socrate lui-même⁸ mais sans le conduire à révoquer la pensée pure, ni à en limiter son usage. Au contraire, la suite du dialogue se poursuit avec une nouvelle définition de la science : celle de

Pour corriger le modèle, Théétète propose alors une seconde version (*Théétète* 199d-200a). Dans celle-ci, les oiseaux peuvent transporter soit des savoirs, soit des ignorances. L'erreur peut alors provenir soit de la confusion entre plusieurs savoirs, soit de la prise en considération d'ignorances. C'est cette deuxième version que j'examine ici, même si, par simplicité, j'avais présenté la première version dans la section ??.

7. *Théétète*, 196a.

8. *Théétète*, 199b-200c.

« l'opinion vraie munie de raison⁹ » La Raison est donc tout de suite posée comme la pierre de touche de la pensée pure, comme le critère fiable qui doit prémunir de l'erreur et de la régression infinie. Or le point de vue de Kant consiste précisément à faire sortir l'opinion vraie de la raison, et c'est en cela que son geste est inaugural.

À ce stade, un lecteur scrupuleux me fera peut-être remarquer que l'objection du colombier ne concerne pas la théorie de Platon, mais probablement quelque théorie défendue par ses contemporains. Le dialogue, dira-t-on, loin de vouloir asseoir cette représentation de la pensée pure, consistait au contraire à en indiquer les impasses. Cette observation est tout à fait légitime. C'est pourquoi j'aimerais, pour prévenir toute hésitation en la matière, montrer que des difficultés analogues surviennent dans la théorie des Idées.

Pour s'en rendre compte, il suffit de poser la question suivante : existe-t-il, quelque part aux confins du monde intelligible, une Idée du Faux ou une idée de l'Erreur ? Si l'on répond par la négative, cela veut dire que la seule façon, pour une pensée pure, de n'être pas vraie, c'est de ne participer d'aucune Idée. Or une pensée pure qui ne participe d'aucune Idée, dans le langage de Platon, ce n'est pas une pensée du tout. L'erreur dès lors ne peut pas être dans la pensée pure, mais uniquement dans le sensible. Cette opinion nous met en peine d'expliquer que des fautes puissent être commises dans le raisonnement abstrait, comme le raisonnement mathématique.

Si l'on répond par la positive, et que l'on admet une idée de l'Erreur, cela ne nous conduit pas à une conclusion plus satisfaisante. Car s'il existe effectivement une Idée de l'Erreur, tout ce qui participe de cette idée est à la fois vrai (puisque se rapportant à une Idée) et à la fois faux (puisque cette Idée est celle de l'Erreur)¹⁰. Les notions de vrai et de faux, si elles existent, doivent donc se dire en plusieurs sens. Ce problème évoque à plusieurs égards certains résultats de la logique contemporaine¹¹ et n'est peut-être que la prémisse des nombreux paradoxes de l'autoréférence.

Qu'est-ce qui est à l'origine de ce type de paradoxes aux yeux de Kant ? Tant que la connaissance portait sur des objets sensibles, l'on disposait d'un critère de vérification, d'une expérience à l'aune de laquelle contrôler l'adéquation en question ; mais sitôt que l'on fait le saut du côté de la pensée pure, la raison n'a plus d'autre fondement qu'elle-même et l'adéquation ne saurait plus être pensée sans s'empêtrer aussitôt dans la contradiction. C'est ce point-là, fondamentalement, qui représente la plus grande critique de Kant,

9. *Théétète*, 201d.

10. Des objections analogues sont avancées par Aristote dans le *ΠΕΡΙ ΙΔΕΩΝ* (ou ce qui en a été conservé grâce aux commentateurs tels qu'Alexandre Aphrodise, et notamment le fait qu'il devrait exister une Idée du non-Être. Pour une discussion plus complète de ces arguments, voir par exemple MANSION, *La Critique de la Théorie des Idées dans le ΠΕΡΙ ΙΔΕΩΝ*

11. Je pense notamment aux théorèmes de Tarski. Le chapitre intitulé *Crise des fondements mathématiques* étudiera cette question plus en détail.

aussi bien à l'adresse de Platon que de toutes les métaphysiques qui l'ont suivi. Loin cependant de se contenter de récuser comme un juge impartial le projet ontologique lui-même, il entreprend de le redresser. Comment ? En limitant l'usage de l'Entendement aux objets d'une expérience possible.

D'aucuns pourraient penser ici que le rejet de la pensée pure signe l'abandon des mathématiques. Ce serait méconnaître l'originalité de la pensée kantienne. Contrairement à Socrate, qui tenait la proposition " $5 + 7 = 12$ " pour un jugement de la pensée pure^[12], Kant y voit un jugement de l'Entendement basé sur l'intuition^[13] :

On est sans doute tenté de croire d'abord que cette proposition $7+5=12$ est une proposition purement analytique, qui résulte, suivant le principe de contradiction, du concept de la somme de sept et de cinq. Mais, quand on y regarde de plus près, on trouve que le concept de la somme de 7 et de 5 ne contient rien de plus que la réunion de deux nombres en un seul, et qu'elle ne nous fait nullement connaître quel est ce nombre unique qui contient les deux autres. [...] Il faut que je sorte de ces concepts en ayant recours à l'intuition qui correspond à l'un des deux, comme par exemple à celle des cinq doigts de la main, ou (comme l'enseigne Segner en son arithmétique) à celle de cinq points, et que j'ajoute ainsi peu à peu au concept de sept les cinq unités données dans l'intuition^[14]. »

Quelle est exactement la nature de cette intuition à laquelle Kant fait référence ici ? Il s'agit d'une intuition pure de l'espace et du temps inhérente à la structure même de la subjectivité, qui existe indépendamment de toute expérience sensible. La position de Kant inaugure l'intuitionnisme mathématique et occupe une place intermédiaire entre le réalisme mathématique de Platon et un empirisme^[15]. À ses yeux, les vérités mathématiques ne ren-

12. Dans le *Théétète* (196a), Socrate soulève la question de savoir d'où provient l'erreur de calcul du type " $5 + 7 = 11$ " : d'une erreur dans la pensée pure ou d'un désaccord avec l'expérience sensible ? Plusieurs pages plus loin, il déclare : « Il faut donc montrer que l'opinion fautive est tout autre chose qu'un désaccord de la pensée et de la sensation. Si en effet c'était cela, nous ne nous tromperions jamais dans nos pensées pures. » (*Théétète*, 196c). Ceci indique que les propositions arithmétiques, pour Platon, ne sont que des jugements de la pensée pure.

13. Le fait que Kant exemplifie le jugement synthétique *a priori* à l'aide de cette proposition " $5 + 7 = 12$ ", que l'on trouve comme telle dans le *Théétète* (*Théétète*, 196a), montre une fois de plus l'étroit dialogue qu'il entretient avec Platon tout au long de sa critique.

14. KANT, *op. cit.*, Introduction V, p. 59-60.

15. Kant se défend d'ailleurs à de nombreux endroits aussi bien de l'idéalisme que de l'empirisme. Voir notamment la section *Réfutation de l'idéalisme* dans l'analytique transcendantale (KANT, *Critique de la raison pure*, trad. TISSOT, pp. 238 - 248), ou encore la *Réfutation de l'argument de Mendelssohn en faveur de la permanence de l'âme*, qui concerne l'idéalisme psychologique (*Dialectique transcendantale*, pp. 340 - 348). Pour l'empirisme, voir aussi la discussion sur Locke et Hume dans le *Passage à la déduction transcendantale des catégories*, (*Analytique transcendantale*, pp. 134 - 136).

voient pas à des objets intelligibles, existant indépendamment du sujet dans le royaume des idéalités. Elles ne sont pas non plus le pur produit d'une construction de l'esprit ; ni davantage une connaissance empirique issue de l'expérience du monde sensible. Elles sont la structure *a priori* de toute expérience possible, ou en d'autres termes, elles sont les formes mêmes de tout raisonnement portant sur les objets sensibles.

2.2.2 La Connaissance *a priori*

Cette refonte de la métaphysique est encore motivée par un autre aspect : celui de rendre possible une connaissance *a priori*. Explicitons-nous. Chez Platon, les Idées sont les objets de la pensée par excellence. Ces Idées existent indépendamment de l'esprit humain. Elles doivent donc déjà être données pour pouvoir être contemplées. Dans ces circonstances, il est difficile de soutenir la possibilité d'une connaissance strictement *a priori*, comme celle des mathématiques. C'est pourquoi la véritable méthode en philosophie spéculative, selon Kant, consiste à poser que ce sont les objets qui se règlent sur la connaissance humaine, et non l'inverse.

Ce renversement de conception est une véritable révolution copernicienne : « Si l'intuition se réglait nécessairement sur la nature des objets, je ne vois pas comment on en pourrait savoir quelque chose *a priori* ; que si, au contraire, l'objet (comme objet des sens) se règle sur la nature de notre faculté intuitive, je puis très bien alors m'expliquer cette possibilité¹⁶. » Tout ce qui fait l'objet de connaissances *a priori* se trouve toujours déjà en nous, préalablement à toute expérience sensible ou intellectuelle. Avec Kant, le jugement synthétique *noétique* de Platon est remplacé par le jugement synthétique *a priori*.

Une conséquence majeure de cette méthode, c'est qu'il n'est plus possible de connaître les choses autrement que comme phénomènes. Dans la mesure, en effet, où ce sont les objets qui se règlent sur la nature de nos concepts, et non l'inverse, nous ne connaissons jamais les objets que dans la modalité de notre perception. Dès lors, il est impossible d'étendre le savoir sur d'autres choses que des objets d'une expérience possible. Or il existe chez l'homme une disposition naturelle, dans le processus de développement continu de sa connaissance, à poser des questions qui excèdent toute représentation de l'intuition sensible. Ces questions sont celles touchant à Dieu, au monde et à l'âme. Le monde est-il fini ou infini ? A-t-il un commencement dans le temps ? A-t-il été créé par un Être Suprême ? L'âme est-elle une substance simple et incorruptible ? Ce sont ces interrogations qui ont nourri tous les systèmes de croyances, qu'ils soient métaphysiques ou cosmogoniques. Dans la mesure où de telles idées ne se réfèrent à aucune expérience possible, elles conduisent toujours à des raisonnements paradoxaux et, pour cette raison,

16. KANT, *Critique de la raison pure*, trad. de BARNI, Préface de la seconde édition, p. 24.

doivent être délaissées. C'est l'objet de la *dialectique transcendante* que de discuter précautionneusement chacune des conclusions logiques où mènent ces considérations, pour montrer leur situation aporétique inextricable.

À bien considérer les choses, le premier résultat de cette critique à laquelle s'attelle le philosophe est somme toute *négative* : il s'agit de montrer que l'usage de la raison pure ne peut pas mener à une connaissance proprement scientifique. Kant proscrie donc la *noësis* de Platon, conçue comme la recherche de l'anhypothétique. Cette restriction, cependant, loin de détrôner la dialectique du piédestal où la philosophie de Platon l'avait hissée, contribue tout au contraire à asseoir sa position. Certes il lui faut descendre quelques marches, certes il lui faut abandonner ses prétentions à embrasser d'un seul regard la totalité du savoir, mais l'exercice de la dialectique en tant que tel est sauvegardé : « il y a toujours eu dans le monde et il y aura toujours une métaphysique, mais toujours aussi on verra s'élever à côté d'elle une dialectique de la raison pure, car celle-ci lui est naturelle. La première et la plus importante affaire de la philosophie est donc d'enlever une fois pour toutes à cette dialectique toute pernicieuse influence en détruisant la source des erreurs¹⁷ »

La Raison inaugurée par Platon, quoique réduite à plus de modestie, survit elle aussi à la critique. Tout d'abord, un certain usage de l'intelligence humaine, qui consiste à connaître de façon *a priori* les objets d'une connaissance possible, est sauvegardé et consolidé. Cet usage repose d'un côté sur les intuitions sensibles, qui donnent au sujet la matière de la connaissance, et d'un autre côté sur une forme de raisonnement discursif, qui permet de catégoriser les données sensibles. Il fait le pont entre la sensibilité et l'intelligible, sans s'égarer dans la pensée pure. Il est donc assez proche de la raison discursive (*dianoia*) dont parlait Platon et peut-être peut-on y voir comme le prolongement ? Kant, et avant lui toute une tradition métaphysique, désignent cette faculté sous le nom "d'entendement". Son royaume est celui des jugements synthétiques *a priori*, qui englobent l'ensemble des mathématiques et de la physique pures.

De plus, les circonscriptions imposées à la raison pure ne s'étendent pas à la raison pratique. Tout au contraire, Kant se montre soucieux de conserver l'ordre moral. Bien que les questions de la liberté, de l'existence de Dieu et de l'immortalité de l'âme, en tant que réflexions spéculatives, se sont évaporées au contact de la critique, la question de la loi morale, pour sa part, a survécu à l'épreuve du feu. Mais peut-être demandera-t-on comment un tel prodige est envisageable ? Il peut paraître surprenant de bâtir une loi universelle *a priori* de la moralité lorsque la liberté humaine ne peut pas être sanctionnée par la Raison de façon confortable. C'est que du brasier de la critique restent les cendres de la croyance. La liberté, certes, est aporétique, mais rien n'empêche l'homme d'y croire. Précisons ce point de vue. Dans la

17. KANT, *op. cit.*, Préface de la seconde édition, p. 36

mesure où la raison ne peut s'exercer que sur les phénomènes, et non sur les choses en soi, il n'y a rien de contradictoire à penser que les noumènes soient libres, tout en maintenant que l'âme, telle qu'elle nous apparaît, ne peut justement apparaître que sous le mode d'une causalité naturelle. Ce point de vue permet de concilier l'antinomie où nous placent, d'une part, la certitude subjective d'être un sujet autonome et, d'autre part, la connaissance de n'être qu'un organisme répondant de lois physico-chimiques immuables.

Ceci fait dire à Kant que les meilleures preuves qu'il faut fournir à l'homme pour l'engager sur le chemin de la foi et de la vertu, ce ne sont pas les preuves abscones des scolastiques, mais les témoignages séduisants de l'expérience humaine¹⁸. Il est vain, et illusoire, de chercher à prouver par voie apodictique les vérités métaphysiques desquelles dépend prétendument l'ordre moral. Il suffit de s'adresser au cœur de l'homme. Qu'il doute de l'existence de Dieu ? Montrez-lui du doigt l'harmonie et la beauté immarcescible de la nature. Qu'il doute de l'immortalité de l'âme ? Attendez qu'il fasse l'expérience que rien de temporel ne saurait jamais le satisfaire pleinement. Qu'il doute de la liberté de sa volonté ? Opposez-lui clairement ses devoirs à ses autres penchants pour faire naître en lui la conscience de ses choix. Ce que menace la critique kantienne, ce ne sont pas les principes moraux et religieux eux-mêmes, sinon le monopole des écoles de pensée qui professent ces principes de façon dogmatique.

Notons que le procès qui consiste à disqualifier la *noësis* au profit de la *dianoia* est d'une importance capitale pour comprendre la signification qu'a prise le concept de Raison dans notre monde occidental, et en particulier à l'apogée du *XX^{ème}* siècle. La *dianoia* opère en choisissant un ensemble d'hypothèses, en décomposant les concepts en éléments plus petits, en calculant mécaniquement les résultats. Elle répond d'une économie de la pensée qui est celle de la logique discursive, de l'ingénierie analytique. Aussi n'est-ce pas étonnant que la Raison de Platon, qui comprenait conjointement une dimension discursive et contemplative, soit devenue la rationalité instrumentale au sens de Weber¹⁹.

Ce n'est d'ailleurs pas un hasard si les critiques de cette rationalité instrumentale, telles qu'on les retrouve dans le postmodernisme ou en phénoménologie, procèdent bien souvent à une revalorisation de la dimension non discursive de la raison. Ceci peut être fait, et l'a souvent été, de deux manières : soit en procédant à une réhabilitation de la *noësis*, c'est-à-dire à un retour aux fonctions intuitives ou contemplatives de la raison ; soit au moyen d'une réhabilitation de la *raison pratique*, passant par la prise en compte des normes, des systèmes de valeurs et du sens dans l'orientation et l'organisation de la rationalité²⁰. Je donnerai deux exemples pour étayer

18. *Ibidem*, p. 37.

19. J'introduis cette idée à titre d'hypothèse. Une analyse plus poussée nous entraînerait en dehors du cadre du présent travail.

20. Il s'agit, ici encore, d'une hypothèse personnelle.

cette idée, tous deux issus de l'école de Francfort²¹.

Le premier exemple, qui illustre selon moi un retour à une forme de *noësis* — mais une *noësis* préalablement débarrassée de ses connotations métaphysiques — peut se trouver dans le concept de *mimesis* développé par Adorno et Horkheimer²². Selon ces auteurs, la rationalité contemporaine s'est élaborée dans une logique de l'identité, consistant d'une part à sortir du discours poétique, c'est-à-dire à désolidariser le signe de la chose ; et d'autre part à subsumer la diversité de l'autre sous des concepts universels. Comme conséquence directe s'est réalisée une double domination : celle de la nature extérieure, objectivée et réduite à un ensemble de ressources exploitables, et celle de la nature intérieure de l'homme, dont la créativité est réprimée. En réponse à cette situation, ils invoquent un retour à une forme de *mimesis*, consistant non plus à se réappropriier la nature pour la rendre à son image, mais à faire place, en soi, à son altérité. Il s'agit donc, pour la subjectivité, de ne plus identifier le non identique à soi, mais de l'identifier en lui-même.

Le deuxième exemple est celui du monde vécu, tel que conçu par Habermas. Si le monde contemporain est dominé par une rationalisation fonctionnelle-instrumentale, c'est à cause de ce qu'il appelle une rationalisation sélective. Or une autre forme de rationalité existe : la raison communicationnelle. Elle se définit par une sortie de la logique stratégique. Au lieu de chercher à exercer sur l'autre une influence, de l'utiliser comme un moyen pour parvenir à ses fins, il s'agit de confronter son point de vue au sien selon la loi du meilleur argument, jusqu'à trouver un consensus. Une telle rationalité, en ce qu'elle s'exerce par le langage au sein d'une collectivité, en ce qu'elle n'est possible que dans des processus d'interaction, est aussi par essence une forme d'agir. Habermas parle d'agir communicationnel. Cela dit, il ne s'agit pas de faire choir toutes les sphères de l'activité humaine sous l'impératif du consensus, sans quoi la société dans son entier serait mise en demeure de dire à chaque instant la raison de son mode d'être ; elle perdrait en efficience, en cohésion et en intégration. Tout au contraire, la raison instrumentale et fonctionnelle doit continuer de régir la société conçue comme un système, mais elle doit laisser le monde des normes, des significations et des valeurs à la raison communicationnelle. Ce monde, c'est le monde vécu.

Maintenant que la critique kantienne de Platon est, nous l'espérons, plus claire à l'esprit du lecteur, il est temps de détailler le contenu de la *Critique de la raison pure*, de façon à mieux comprendre le statut du jugement synthétique *a priori*, et en particulier la nature de la synthèse dont il est question ici.

21. L'École de Francfort est un mouvement de philosophie sociale influencé par le marxisme et l'idéal des lumières, qui se donne pour tâche de critiquer le capitalisme afin de promouvoir l'émancipation des individus. On y trouve la figure de Max Horkheimer (1895-1973), de Theodor W. Adorno (1903-1969) ou, plus récemment, d'Axel Honneth (né en 1949).

22. Voir RICARD [?], *Mimésis et vérité dans l'esthétique d'Adorno*.

2.2.3 La Révolution copernicienne

Aux yeux de Kant, l'histoire des sciences a doté les mathématiques et la physique de méthodes appropriées à leurs objets, dont il n'est pas possible d'adjoindre ou d'ôter quoi que ce soit²³. En ce qui concerne la méthode mathématique, elle fut trouvée au moment où est née la géométrie euclidienne, au moment où s'est imposée la conviction selon laquelle les propositions mathématiques peuvent faire l'objet de démonstrations apodictiques. Qu'entendons-nous par là ?

La grande idée des Grecs, ce qui leur a permis de s'émanciper des mathématiques égyptiennes ou sumériennes, est d'avoir pensé les propriétés des figures géométriques sans éprouver la nécessité de les représenter dans l'espace²⁴. Ce que la science dit du triangle, elle le dit de l'essence du triangle, c'est-à-dire de tout triangle, quel qu'il soit. Ce faisant, ses jugements sont parfaitement universels et *a priori*. La géométrie euclidienne a acquis une telle perfection qu'il n'est pas possible de la faire progresser davantage. Il est tout au plus permis d'en exposer plus clairement les résultats, ou de l'inclure dans une théorie plus générale au titre de cas particulier, à l'image de la position actuellement occupée par la géométrie euclidienne au sein des géométries non euclidiennes.

Dans le cas de la physique, la méthode fut plus pénible à paraître à la lumière du jour. Et pour cause, son objet est intimement mêlé d'expérience sensible, si bien qu'il a fallu déterminer comment les principes de la raison devaient s'accommoder des observations de la nature, et inversement. C'est à Bacon, à qui il dédie d'ailleurs une épithète²⁵, que Kant attribue la paternité de la juste méthode. Cette dernière se faisait déjà entrevoir, comme par un jeu de reflets, dans les expériences de ses prédécesseurs : les plans inclinés de Galilée, les baromètres de Torricelli et les théories controversées de Sthal. La raison apprend progressivement à reconnaître et à orienter son intervention dans l'expérience. Ce qu'elle entrevoit d'un dispositif expérimental n'est jamais rien d'autre que ce qu'elle y a mis elle-même d'après ses propres plans. Loin de recueillir la parole de la nature à la façon d'un écolier

23. KANT, *op. cit.*, Préface de la seconde édition, pp. 19-22.

24. « Le premier qui démontra le triangle isocèle (qu'il s'appelât Thalès ou de tout autre nom) fut frappé d'une grande lumière ; car il trouva qu'il ne devait pas s'attacher à ce qu'il voyait dans la figure, ou même au simple concept qu'il en avait, mais qu'il n'avait qu'à dégager ce que lui-même y faisait entrer par la pensée et construisait *a priori*, et que, pour connaître certainement une chose *a priori*, il ne devait attribuer à cette chose que ce qui dérivait nécessairement de ce qu'il y avait mis lui-même, suivant le concept qu'il s'en était fait. » (KANT, *ibidem*, p.20).

25. « La physique arriva beaucoup plus lentement à trouver la grande route de la science ; car il n'y a guère plus d'un siècle et demi qu'un grand esprit, Bacon de Verulam, a en partie provoqué, et en partie, car on était déjà sur la trace, stimulé cette découverte, qui ne peut s'expliquer que par une révolution subite de la pensée. Je ne veux ici considérer la physique qu'autant qu'elle est fondée sur des principes empiriques. » (KANT, *op. cit.*, Préface de la seconde édition, p. 21).

s'abreuvant de l'enseignement de son maître, la raison doit se comporter comme un juge qui dispose du droit d'interroger, de contraindre ses témoins si nécessaire. Au travers des expériences, la nature est mise en demeure de répondre de ce qu'elle est. Il y a une étroite relation qui unit les principes de la théorie aux dessins du dispositif expérimental.

Contrairement aux cas de la physique et des mathématiques, il n'est jamais arrivé, dit Kant, que la métaphysique accomplisse un réel progrès méthodologique, car elle « n'a pas encore été assez favorisée du sort pour entrer dans le sûr chemin de la science²⁶ ». La raison de cette lacune tient selon Kant dans deux causes. Tout d'abord, la connaissance *a priori* ne peut rien attribuer d'autre aux objets que ce qu'elle y a mis elle-même, si bien que les recherches doivent être orientées sur la raison en tant que faculté, et non sur ses objets spéculatifs. D'autre part, la raison forme une telle unité que tous ses principes sont solidaires les uns des autres. Il faut les analyser chacun sous tous leurs rapports pour vérifier la cohérence de l'ensemble.

Ces deux obstacles une fois franchis, rien ne saurait empêcher la métaphysique de devenir une science. Contrairement aux mathématiques et à la physique, dont les objets se tiennent au-dehors de la connaissance même, il est possible pour la métaphysique d'embrasser d'un coup l'ensemble de ce qu'il y a à savoir. Sitôt qu'elle aura été conduite sur la bonne voie, et avec la bonne méthode, il n'y a pas lieu de croire qu'elle ne puisse pas, comme la logique d'Aristote avant elle, devenir un savoir complet et achevé. Elle pourra, aux yeux de Kant, être transmise en l'état aux générations futures, les seuls changements nécessaires touchant, non à son contenu, mais à une meilleure exposition de celui-ci.

2.3 La Sensibilité : l'Homme qui sent

La grande idée kantienne, le point de départ de toute son interrogation, consiste à voir les résultats de la science moderne, dont la physique newtonienne et la géométrie euclidienne représentent les plus éminents archétypes, comme un ensemble de propositions synthétiques *a priori*²⁷. Bien entendu, ce point peut être contesté, et l'a été dans l'histoire récente des sciences physiques et mathématiques. Les géométries non euclidiennes, en particulier, ont explicitement soulevé la question de savoir si des vérités telles que « La droite est la plus courte distance entre deux points » consistaient comme le prétend Kant en des jugements synthétiques *a priori*²⁸.

Cela dit, admettre comme le fait Kant que la connaissance scientifique ne se développe ni uniquement sur des jugements analytiques ni uniquement

26. *Ibidem*, pp. 22-23.

27. *Ibidem*, Introduction V, pp. 58-62.

28. Cela fera l'objet du prochain chapitre : *La crise des fondements des mathématiques*. Voir en particulier *Objections à Kant*, section 4.4

sur des jugements *a posteriori*, c'est poser la question de savoir comment peut être opérée de façon apriorique la production de jugements scientifiquement valables ? En particulier, Kant demande quel doit être l'élément X permettant de mettre en relation les deux termes intervenant dans un jugement synthétique à vocation universelle, si ce n'est pas l'expérience ? Ceci le conduit à rechercher les conditions de possibilité de l'expérience scientifique elle-même.

Illustrons cette démarche à l'aide d'un exemple. Le principe suivant : « Tout effet est le produit d'une cause » est un jugement synthétique *a priori* utilisé par la science, dont la vérité ne dérive pas de l'expérience. La démarche empirique, en effet, ne conduirait in fine qu'au résultat humien selon lequel : « Dans toutes les expériences réalisées jusqu'à présent, les effets sont les produits d'une cause », sans jamais affirmer l'universalité du principe. C'est pourquoi un tel principe, selon Kant, doit pouvoir être dérivé, non de l'expérience, mais de la structure même de la connaissance. Il s'agit dès lors, pour Kant, d'exhiber la structure apriorique de chacune des facultés de la connaissance que sont respectivement la Sensibilité, l'Entendement et la Raison. Cette démarche qui, partant de l'existence de certains types de jugement *a priori*, cherche à déterminer leurs conditions de possibilité, est dite *transcendantale*.

La sensibilité, selon Kant, est structurée par deux intuitions, l'intuition pure de l'espace et l'intuition pure du temps, qui correspondent respectivement au sens externe (la perception du monde) et au sens interne (la perception de soi-même). Ces intuitions sont ce qui rend l'expérience possible, en ce sens précis qu'aucun phénomène ne peut être pensé en dehors du temps et de l'espace. « Ce n'est pas la sensation », dit Kant, « qui me donne la représentation de l'espace, puisque, sans cette représentation même, je ne saurais me représenter les choses de cette manière. Il faut donc qu'elle existe déjà en moi²⁹. » Ou encore : « Il est impossible de se représenter qu'il n'y ait point d'espace, quoique l'on puisse bien se représenter qu'il n'y ait point d'objet³⁰. » Il en va de même avec le temps. « On ne saurait supprimer le temps lui-même par rapport aux phénomènes en général, quoique l'on puisse bien les retrancher du temps par la pensée. Le temps est donc donné *a priori*. Sans lui, toute réalité des phénomènes est impossible³¹. » Si la matière de la sensibilité est donc bien donnée par la sensation de façon *a posteriori*, sa forme en revanche est présente dans l'esprit de façon apriorique.

Il s'ensuit deux conséquences remarquables. Tout d'abord, les choses ne nous sont pas accessibles dans l'absolu, telles qu'elles sont en elles-mêmes, mais seulement au travers du prisme de ces intuitions, c'est-à-dire, de notre propre subjectivité. En d'autres termes, les choses ne sont connaissables

29. KANT, *op. cit.*, Esthétique transcendantale, De l'espace, §2, p. 77.

30. *Ibidem*, p. 78.

31. *Ibidem*, Esthétique transcendantale, Du temps, §4, p. 86.

qu'en tant que phénomènes : le monde nouménal est hors de portée de la connaissance sensible. D'autre part, le temps et l'espace n'ont pas d'existence absolue. Dans l'exacte mesure où le temps et l'espace sont des conditions *a priori* de la sensibilité, ils ne sont rien en dehors des représentations du sujet. Ce ne sont pas des réalités à part entière, pas plus que des propriétés inhérentes aux objets de l'expérience. Ces intuitions ont certes une valeur objective, en ce sens qu'elles s'appliquent à tous les objets possibles de l'expérience, mais non une valeur absolue. Pour le dire dans le langage kantien, elles sont à la fois des réalités empiriques et des idéalités transcendantales :

« Notre examen de l'espace nous en montre donc la réalité (c'est-à-dire la valeur objective) au point de vue de la perception des choses comme objets extérieurs ; mais il nous en révèle aussi l'idéalité au point de vue de la raison considérant les choses en elles-mêmes, c'est-à-dire abstraction faite de la constitution de notre sensibilité³². »

Notons pour finir que l'espace et le temps ne peuvent pas faire l'objet de la pensée pure, mais qu'ils ne se donnent à voir qu'au travers d'une représentation sensible. Au même titre qu'il est impossible de penser une chose en dehors de l'espace et du temps, il est tout aussi impossible de penser le temps et l'espace indépendamment des choses de l'expérience. Cette observation constitue une première limite que le sujet rencontre lorsqu'il veut penser son intériorité. Le sujet ne peut se saisir, au moyen du sens intérieur, que tel qu'il s'apparaît à lui-même ; c'est-à-dire non tel qu'il est véritablement en soi, mais seulement en tant que phénomène. Il n'a concernant lui-même qu'une connaissance empirique. Par suite, une différenciation apparaît immédiatement entre le sujet perceptif, tel qu'il est en son intériorité, et la représentation qu'il acquiert de cette intériorité. Or dès lors que le vrai soi n'est pas accessible à la connaissance, il devient pertinent de se demander comment est rendue possible une connaissance réflexive du sujet et de ses facultés ? N'est-ce pas là précisément ce que Kant entend faire dans la critique de la raison pure ? Cette éventuelle circularité de la pensée kantienne ne doit pas être discutée ici, mais je la présente pour bien faire remarquer que le problème de l'autoréférence qui avait été repéré dans la théorie de Platon n'est peut-être que déplacé avec la solution kantienne.

32. KANT, Esthétique transcendantale, De l'espace, §4 Conséquences tirées de ce qui précède, B, p. 83. Plus loin, Kant fait une remarque analogue sur le temps : « Nous n'admettons donc pas que le temps puisse prétendre à une réalité absolue, comme si, même abstraction faite de la forme de notre intuition sensible, il appartenait absolument aux choses à titre de condition ou de propriété. » (*Ibidem*, Esthétique transcendantale, Du temps, §6 Conséquences tirées de ce qui précède, C, p. 91)

2.4 L'Entendement : L'Homme qui pense

Si la Sensibilité est la faculté relative à la sensation, qui rend possibles les représentations, l'Entendement pour sa part est la faculté de juger. Or qu'est-ce que juger ? C'est procéder à la synthèse des différentes représentations de la sensibilité de manière à constituer la connaissance d'un objet déterminé : comme lorsque je dis « Ceci est un métal » ou encore « Tout métal est un corps³³. » Les intuitions pures de l'espace et du temps ne suffisent donc pas à la connaissance. C'est uniquement grâce à l'action de l'Entendement que le divers de la sensibilité peut être unifié sous un concept. Pour résumer cela en une phrase lapidaire : sans la sensibilité rien ne serait *donné*, mais sans l'Entendement rien ne serait *pensé*.

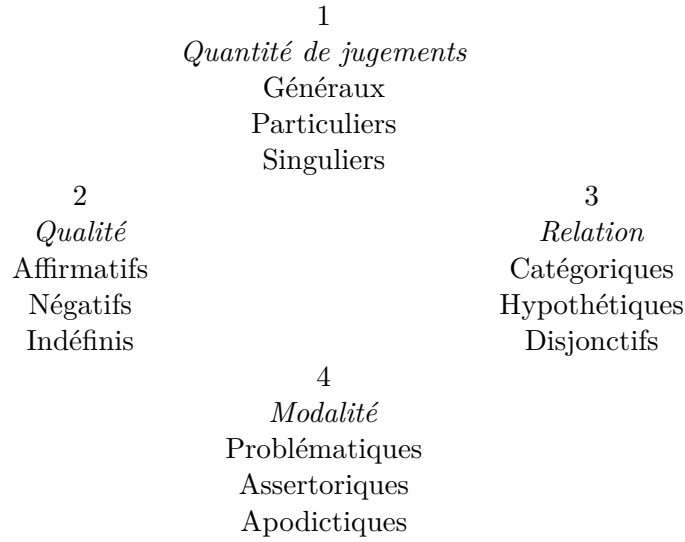
Kant distingue la logique générale, dont le rôle est de définir les formes pures de la pensée indifféremment de son origine, ou de son rapport à un objet, de la logique transcendantale, qui pour sa part vise à délimiter la valeur objective d'un concept *a priori* de l'Entendement. L'on pourrait dire de la logique générale qu'elle est purement formelle : elle étudie comment les concepts s'enchaînent en suivant les lois de l'Entendement, sans s'interroger sur l'origine empirique ou intuitive de ces concepts et de ces lois.

Pour sa part, la logique transcendantale pose la question de savoir ce qui garantit la validité du jugement portant sur un objet, de façon apriorique. Kant nous met toutefois en garde de ne pas y mêler des considérations d'ordre psychologique, mais de rester bien précautionneusement dans une pure théorie de la connaissance. Il n'est pas question en logique transcendantale de se demander quelles conditions empiriques doivent être satisfaites pour produire un jugement valable, comme l'attention, le dénombrement, etc. mais uniquement quelles sont les structures universelles de l'Entendement qui rendent le jugement possible.

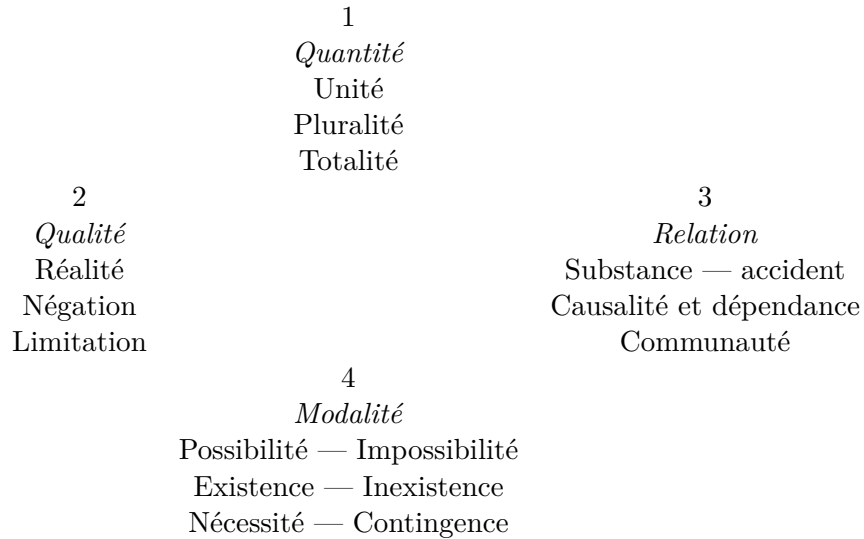
La logique générale sert de point de départ à la logique transcendantale. Pour déterminer en effet la structure apriorique de l'Entendement, Kant commence d'abord par exhiber les fonctions logiques présidant aux jugements. Ces fonctions sont de simples opérations de pensée : les jugements catégoriques, hypothétiques, disjonctifs, etc. et s'apparentent à une logique modale du premier ordre³⁴ :

33. *Ibidem*, Analytique des concepts, Chapitre I, section 1, p. 128.

34. *Ibidem*, Analytique transcendantale, Chapitre I, section 2, §9, p. 128.



Ensuite, Kant applique ces fonctions aux intuitions pures de l'espace et du temps et obtient ainsi les catégories de l'Entendement, qui ne sont rien d'autre que les concepts *a priori* présidant à l'unification du divers de la sensibilité. Parmi ces principes, l'on dénombre ceux de la quantité, de la qualité, de la relation et, enfin, de la modalité³⁵:



2.4.1 Kant vs François Chollet

Pour donner un éclairage pratique à ces catégories de l'Entendement, peut-être n'est-il pas inutile de tisser un parallélisme entre la théorie de Kant et les quatre aptitudes de l'intelligence artificielle générale identifiées par

35. *Ibidem*, Analytique transcendantale, Chapitre I, section 3, §10, p. 138.

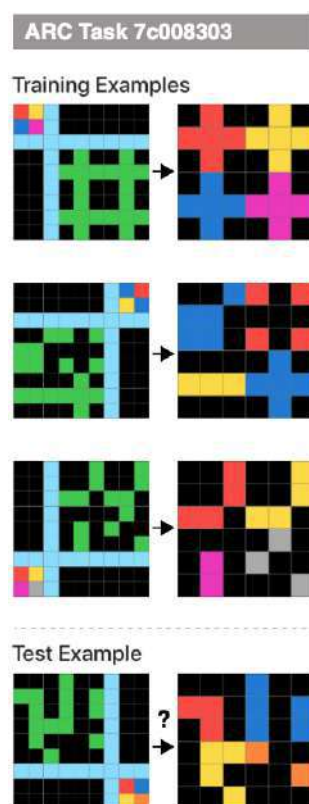


FIGURE 2.5

François Chollet dans le chapitre ? Pour rappel, François Chollet cherchait à déterminer quelles opérations cognitives étaient nécessaires pour qu'une machine réussisse les exercices ARC, une sorte de test de QI artificiel défini par un ensemble de problèmes de généralisations similaires aux matrices de Raven (voir figure 2.5). Il distinguait à ce titre quatre ingrédients phares : (1) la persistance de l'objet, (2) la géométrie / topologie, (3) l'arithmétique et (4) l'orientation vers un but. Je vais rappeler les définitions de ces quatre aptitudes et montrer en quoi elles s'apparentent, ou se différencient au contraire, des catégories de l'Entendement de la théorie kantienne.

L'arithmétique est la faculté à trier les objets en fonction de leur forme, de leur couleur, de leur apparence ou de leurs mouvements en s'appuyant sur l'addition, la soustraction et la comparaison. Elle implique notamment trois opérations de pensée : (1) la capacité à distinguer l'unité (un pixel est vert), (2) la capacité à distinguer la pluralité (il existe plusieurs pixels verts) et (3) la capacité à considérer la totalité (tous les pixels sont verts). Ces opérations consistent donc à abstraire un objet particulier au sein d'un tout, en le comparant aux autres éléments de ce tout, et en constatant quelles sont les caractéristiques qui le distinguent. Ces comparaisons correspondent

justement, dans la critique kantienne, aux catégories de la Quantité, qui permettent de poser des jugements universels (« tous les carrés sont verts ou jaunes »), des jugements particuliers (« il existe des carrés jaunes ») et singuliers (« ce carré est jaune »). Il est donc possible de penser que l'arithmétique, en tant que compétence mathématique susceptible d'être entraînée, s'appuie sur une structure innée plus fondamentale de l'Entendement humain. L'usage du mot « innée » ne veut toutefois pas dire que cette structure existe telle quelle dès la naissance, mais qu'il s'agit d'une faculté qui se déploie d'après une trajectoire développementale prédéfinie.

La *persistance de l'objet* est la capacité à repérer des objets et à les manipuler sans qu'un changement d'orientation ou de contexte fasse disparaître l'identification de cet objet. Il s'agit donc, premièrement, de déterminer si un objet existe ou non au sein de l'image, mais également de comprendre s'il existe de façon nécessaire ou accidentelle. Si l'on reprend l'exemple de la figure 2.5, on constatera que la présence des carrés jaunes, au milieu des figures vertes, est nécessaire. En revanche, la présence des carrés verts est contingente. Il n'y a aucune règle concernant leur agencement. Il s'ensuit également qu'il est impossible d'avoir un carré jaune en dehors d'une figure verte fermée. Ces trois modalités, l'existence/la non-existence, la nécessité/la contingence et la possibilité/l'impossibilité, sont justement reprises par la catégorie kantienne de la Modalité. C'est ce qui permet de distinguer les jugements problématiques (il est possible qu'un carré soit vert), les jugements assertoriques (ces carrés sont verts) et le jugement apodictique (un carré n'est jaune que s'il est dans une figure verte fermée).

L'*orientation vers un but* est la capacité à imputer des intentions ou des buts à certains objets. La résolution de la matrice de Raven de l'exemple 2.5 est facilitée si vous pensez que le « but » des pixels jaunes est de remplir les espaces clos. Vous pouvez les voir comme un troupeau qui cherche le refuge d'un enclos à la tombée de la nuit. Cette aptitude ne correspond pas, *stricto sensu*, aux catégories identifiées par Kant. Elles semblent néanmoins appartenir à la catégorie de la Relation. Cette dernière prescrit en effet des rapports de causalité entre les objets. Elle permet de différencier les jugements catégoriques (le troupeau cherche un refuge!), hypothétiques (si le troupeau veut survivre, il doit trouver un enclos) et disjonctifs (soit l'animal trouve un enclos, soit il disparaît). Je reconnais que ce parallélisme que je dresse entre l'orientation vers un but et la catégorie kantienne de la relation ne saurait être que purement analogique, et résiste difficilement à une analyse en règle. La raison en est que Kant ne s'intéresse pas à l'intentionnalité, mais aux différentes formes de jugement — envisagées de façon objective.

Pour finir, *géométrie et la topologie* sont, d'après François Chollet, la capacité à réaliser des manipulations spatiales sur les objets, telles que des rotations, des symétries orthogonales, des translations, des déformations, des répétitions, des associations, etc. Cela demande de systématiser les qualités des objets et les relations existant entre eux. Il faut notamment comprendre

les situations où un changement, dans l'image, est de nature accidentelle ou nécessaire. Lorsqu'une rotation de 45 degrés a lieu, il est par exemple nécessaire que certains points soient déplacés, mais pas de n'importe quelle manière. Cet ingrédient dont je parle se retrouve, chez Kant, à cheval entre la catégorie de la Qualité et la catégorie de la Relation. La Qualité permet d'assigner des propriétés intrinsèques à des objets, comme le fait d'être vert ou d'être jaune. La Relation concerne, pour sa part, non pas des qualités intrinsèques, mais les dispositions que ces objets ont les uns avec les autres : la distance spatiale, la présence de symétrie, etc.

L'analyse que je viens de faire des critères de François Chollet semble suggérer que, pour obtenir une intelligence artificielle forte, il est nécessaire de doter les algorithmes d'une certaine faculté de juger. Cette dernière repose sur un ensemble de structures intrinsèques de la pensée. Autrement dit, nos jugements suivent un ensemble de distinctions conceptuelles telles que la nécessité, la contingence, l'existence, la non-existence, la possibilité, l'impossibilité, ou encore, la causalité. Ces concepts *a priori* sont les briques élémentaires du raisonnement, elles sont ce sans quoi aucune proposition logique ne pourrait être exprimée ou comprise. Il est dès lors très important de se demander quelles sont les conditions pour qu'une machine puisse comprendre et utiliser ces concepts. Je traiterai de cette importante problématique dans le chapitre [3.4](#).

Dans l'immédiat, j'aimerais revenir quelques instants sur la pensée de Kant. Dans la démarche qui est celle de la Critique de la Raison pure, définir les catégories logiques de l'Entendement ne fait pas tout ; il reste encore à procéder à l'analyse de ce qui fait leur validité, tant du point de vue objectif que subjectif. Précisons cette distinction. Objectivement, il convient de fournir la garantie qu'il existe effectivement des objets auxquels les concepts que nous avons définis, et qui nous viennent de la logique pure, s'appliquent. En d'autres termes, il faut montrer que ce ne sont pas des catégories vides.

Du point de vue subjectif, il faut en outre expliquer le fonctionnement interne de l'Entendement ; c'est-à-dire qu'il faut expliquer ce qui rend possible pour la conscience de synthétiser effectivement le donné divers de la sensibilité pour le subsumer sous une catégorie. En fait, cette deuxième question n'est pas indispensable à la critique de la raison proprement dite, mais elle donne un éclairage très précieux. Kant le dit lui-même : « bien que cet examen ait une grande importance relativement à mon but principal, il n'y appartient pourtant pas essentiellement, car la question capitale est toujours de savoir ce que l'Entendement et la Raison, libres de toute expérience, peuvent connaître, et jusqu'à quel point ils peuvent pousser leur connaissance, et non pas comment la faculté même de penser est possible^{[36](#)}. »

36. *Ibidem*, Préface de la première édition, pp. 11-12.

2.4.2 La Dédution objective

Il pourrait se faire qu'aucun contenu sensible ne corresponde aux catégories ainsi déterminées ; que les phénomènes se succèdent de façon tellement chaotique que rien en eux ne fournisse une règle correspondant, par exemple, au principe *a priori* de la cause et de l'effet. Dans ce cas, les phénomènes présenteraient des objets à notre intuition, mais ne seraient pas intelligibles, et il s'ensuivrait que les catégories de l'entendement ne seraient que de pures formes vides. « Expliquer comment des concepts peuvent se rapporter *a priori* à des objets, voilà ce que je nomme la déduction transcendantale de ces concepts³⁷ »

Cette objection, heureusement, s'évapore en fumée sitôt que l'on prend la peine de se rappeler que, pour Kant, ce n'est pas l'entendement qui doit se régler sur les objets, mais les objets qui se règlent à l'entendement. Subtilité et jeu de langage ? Loin s'en faut. Ce que Kant soutient, c'est que l'entendement est, par ses concepts, l'auteur de l'expérience. Les objets, quant à leurs apparitions phénoménales, s'adaptent donc aux concepts *a priori* qui sont en nous. Par suite, il est impossible que ces concepts ne trouvent pas d'objet. Ou autrement dit, ces principes *a priori* sont toujours les principes des phénomènes "véritables", bien que ces phénomènes ne nous enseignent rien sur la chose en soi.

Une importante conséquence qui découle de tout ceci, et qui a pour effet de restreindre considérablement l'usage de l'Entendement, c'est que les catégories se déploient toujours dans une expérience ; sinon une expérience concrète et véritable, tout au moins une expérience *possible*. Le terme "possible" est ici d'une importance capitale. Dans le cas de jugements *a posteriori*, comme ceux portant sur des objets concrets au travers d'un dispositif expérimental effectif, il ne fait aucun doute que les catégories ont une objectivité. Dans le cas des jugements synthétiques *a priori* en revanche — ceux-là mêmes dont Kant tente de prouver l'existence — l'expérience ne peut pas toucher des objets effectifs, présentement donnés à nos sens. Il n'en reste pas moins que les termes du jugement, comme les idées de "corps pesant" ou de "métal", ne peuvent être pensés que dans le cadre d'une expérience possible. Le meilleur exemple de ce type de raisonnement en physique, ce sont les expériences de pensée. Ce sont elles qui servent à éprouver les lois de la nature et à en comprendre la cohérence ; ou au contraire à les invalider. En dehors de cette expérience, les catégories sont sans objet.

Kant insiste lourdement sur cet aspect. Certes les catégories ont une existence apriorique, certes elles permettent de définir des champs de connaissance pure, comme celui des mathématiques ou de la physique. Mais le fait de tenir sa validité de façon apriorique n'empêche pas la connaissance de devoir, pour être rendue effective, être accompagnée d'une expérience possible. Les idées mathématiques n'échappent pas à la règle. Bien qu'il n'existe

37. *Ibidem*, Analytique transcendantale, Chapitre II, section 1, §13, p. 148.

peut-être pas dans le monde un seul objet parfaitement conforme à l'idée du triangle, cette idée n'en tire pas moins son sens de la possibilité d'être appliquée à un contenu sensible.

Notons pour conclure que les lois du jugement, que sont les catégories, sont les conditions de possibilité de l'expérience. Ce n'est pas l'expérience qui rend les concepts possibles, mais c'est bien au contraire les concepts qui rendent l'expérience possible. Néanmoins, le concept doit avoir un contenu, il doit se porter à la rencontre d'un donné qu'est l'intuition sensible. À ce titre, la pensée n'est pas un solipsisme, elle n'est pas une pure immanence ; mais elle renvoie bien au contraire à un au-delà de la pure pensée. Pour Kant, la connaissance réelle est une connaissance synthétique. C'est sans doute ce qui lui fait dire, non sans éloquence, que sa pensée évite aussi bien l'écueil de l'empirisme que de l'idéalisme. Bien que les noumènes soient hors d'accès, bien que le sujet se donne à lui-même ses propres lois phénoménales, les représentations ne sont pas de simples illusions. Il y a véritablement une rencontre avec de l'altérité, avec une matière qui n'est pas contenue de façon apriorique dans le sujet. Cette altérité, cependant, est donnée dans les modalités fixées par les intuitions pures de l'espace et du temps, puis informée par les catégories de l'Entendement.

2.4.3 La Déduction subjective

Kant adresse ici la question de la constitution de la connaissance : comment l'acte de synthèse s'opérant dans l'Entendement est-il possible ? Il prend l'exemple d'un homme qui pénétrerait dans un somptueux palais. Ses sens sont frappés par l'éclat des vitraux, par le charme épuré du mobilier, par la grâce des colonnades, par tout autant de détails qui, chacun, se disputent son attention. Pourtant, sa conscience ne se déchire pas entre toutes ces impressions diverses. Notre homme est capable d'identifier tous ces éléments, de les différencier les uns des autres et de les isoler de leur environnement ; mais il garde dans le même temps la certitude de l'unité de sa propre conscience. C'est bien un même homme qui a conscience du vitrail, du mobilier, de la colonnade, et non une pluralité de consciences hétérogènes. En d'autres termes : la conscience est constituée d'une multiplicité de perceptions sensibles, mais elle reste invariablement une. C'est donc une unité synthétique *a priori*. Bien sûr, cette unité est le quotidien de notre expérience sensible. Elle nous est devenue si naturelle qu'il peut paraître étrange de se demander d'où elle procède. C'est bien pourtant la question fondamentale adressée par Kant, lorsqu'il demande « ce qui contient le principe de l'unité de différents concepts au sein des jugements et par conséquent de l'unité de l'entendement lui-même³⁸ ».

Face à cette question, il fait immédiatement remarquer que la conscience

38. *Ibidem*, Analytique transcendantale, Chapitre II, section 2, §15, p.160.

unifiée de soi-même n'est pas à rechercher dans une catégorie, comme la catégorie de l'unité, ni dans une quelconque catégorie supérieure à celles qui ont déjà été données. En effet, la catégorie ne fait que représenter une fonction logique du jugement. En tant que telle, elle suppose déjà que soit réalisée l'unité du concept, ou en d'autres termes, elle ne peut pas expliquer comment les catégories s'appliquent. Procéder ainsi ne ferait que déplacer la question. Ce serait vouloir rendre compte du fonctionnement d'une machine inconnue à l'aide de cette même machine.

Où faut-il dès lors chercher ce principe qui est à l'origine de l'unité de nos représentations sensibles ? Réponse : dans le « Je pense » cartésien, qui ramène la diversité des perceptions sensibles à une seule et même conscience de soi-même. « Le *je pense* doit pouvoir accompagner toutes mes représentations ; car autrement il y aurait en moi quelque chose de représenté, qui ne pourrait pas être pensé, ce qui revient à dire ou que la représentation serait impossible ou du moins qu'elle ne serait rien pour moi³⁹ »

Précisons cette idée. Les représentations diverses données dans une intuition, par exemple dans celle d'un palais ou d'un temple, ne sont toutes ensemble mes représentations que parce que toutes ensemble elles appartiennent à une même conscience. La conscience n'est donc pas éparpillée dans chacune de ces représentations, mais elle se trouve tout entière présente en chacune d'entre elles. Ce principe est ce que Kant appelle *l'aperception transcendante*. Celui-ci est très important dans le système kantien. Elle est à l'Entendement ce que l'intuition est à la Sensibilité : à savoir, une forme qui rend possible l'exercice de l'Entendement de façon *a priori*. « Le principe suprême de la possibilité de toute intuition, par rapport à la sensibilité, était, d'après l'esthétique transcendante, que tout ce qu'elle contient de divers fut soumis aux conditions formelles de l'espace et du temps. Le principe suprême de cette même possibilité, par rapport à l'entendement, c'est que tout ce qu'il y a de divers dans l'intuition soit soumis aux conditions de l'unité originairement synthétique de l'aperception⁴⁰ »

Cette forme générale à son tour se subdivise en un certain nombre de formes particulières correspondant aux différentes fonctions logiques du jugement, que Kant a appelées les catégories. L'aperception transcendante est donc à l'origine de tous les concepts de l'Entendement, et en représente le principe supérieur de fonctionnement.

2.4.4 Le Schématisme kantien

Tant qu'elles n'ont pas été subsumées sous une catégorie, les données sensibles en tant que telles ne peuvent pas être *perçues* par la conscience. Elles ne sont que des perceptions confuses, sans forme ni unité distinguables. C'est pourquoi, maintenant que les possibilités objectives et subjectives des

39. *Ibidem*, Analytique transcendante, Chapitre II, section 2, §16, p. 160.

40. *Ibidem*, Analytique transcendante, Chapitre II, section 2, §17, p. 164.

catégories ont été vérifiées, il reste encore à préciser comment s'opère la synthèse de façon concrète ; il faut en montrer le mécanisme.

À cette fin, Kant relève la présence de deux ingrédients : l'imagination et l'intuition pure du temps. Pourquoi le temps et non l'espace ? Parce que le temps est commun à tous les phénomènes, aussi bien ceux qui se passent en nous qu'à l'extérieur de nous, tandis que l'espace est seulement une forme *a priori* des phénomènes répondant du sens externe. Or pour que la synthèse s'opère, il est impératif de relier la catégorie, qui est *a priori*, avec ce que le phénomène a lui-même d'*a priori* en lui. L'imagination, quant à elle, a une fonction *hénologique*. Elle produit des schèmes et des images servant de médiation entre l'unité synthétique de la conscience et la multiplicité des représentations sensibles.

Qu'est-ce qu'un schème et qu'est-ce qui le différencie d'une image ? Quels liens entretiennent-ils l'un l'autre ? Le schème n'a pas de composante visuelle. Ce n'est pas une représentation, ni un concept, mais une méthode permettant de construire l'objet sensible dans l'intuition. Expliquons-nous. Le concept de triangle a pour extension une infinité de triangles possibles : les triangles obtusangles, acutangles, isocèles, rectangles, etc. La fonction cognitive nous permettant d'associer l'image d'un triangle à son concept mathématique passe par un mécanisme de reconnaissance mentale consistant à vérifier que cette image du triangle correspond effectivement à la définition que nous en détenons. L'on suit mentalement les contours de la figure géométrique, on la retrace dans sa tête, de façon à s'assurer qu'elle comporte bien trois côtés. Ce processus se déroule dans le temps et il est parfaitement apriorique : même en l'absence d'un "vrai" triangle, il est toujours permis de construire dans l'esprit un triangle. L'existence d'un objet mathématique, selon ce point de vue, est constructive. Existe ce qu'il est possible de définir au travers d'un processus numérique ou géométrique consistant en un nombre fini d'opérations.

Sitôt que l'image du triangle a été identifiée à l'aide du schème, il se forme une image mentale de l'objet. Cette image floue n'est ni assez précise pour correspondre à un vrai triangle sensible, ni assez abstraite pour désigner le concept du triangle dans toute sa généralité. Elle comporte d'ailleurs souvent des propriétés n'appartenant pas à la définition du triangle, comme des couleurs, une certaine épaisseur ou une taille déterminée. C'est néanmoins cette image mentale, issue du schème, qui servira d'intermédiaire entre l'objet sensible et le concept pour permettre la synthèse.

Cette théorie du schème change l'importance donnée à l'imagination. Platon voyait dans l'imagination une faculté dévoyée, et ce à double titre : non seulement elle portait sur le sensible, mais loin de nous fournir la représentation de l'objet lui-même, elle ne nous en faisait voir qu'une pâle copie^[41] ! C'est pour cette raison, sans doute, que l'imagination a été assi-

41. L'imagination peut se trouver dans les dialogues de Platon sous les mots grecs *phan-*

milée au niveau le plus vulgaire de la connaissance : *l'eikasia*. Elle fut si bien méprisée par la tradition philosophique qu'elle était tantôt prise pour une source d'erreurs, tantôt pour une source de tentation. La religion et la métaphysique s'en étaient fait les plus fidèles contempteurs. Or avec Kant, elle est revalorisée. Elle n'est plus seulement la capacité de produire des images, mais elle est aussi et surtout ce qui rend possible le déploiement de l'intuition et la représentation sensible de nos concepts. Ce n'est d'ailleurs pas un hasard s'il lui consacre tout un chapitre dans son *Anthropologie*⁴², et une place importante dans le jugement esthétique⁴³.

En dépit de ce rôle joué par l'imagination, la synthèse reste du côté de l'intelligible. La première raison à cela, et sans doute la plus évidente, c'est la place primordiale qu'y occupent les catégories. Leur fonction consiste à dégager du sensible des unités conceptuelles, en opérant une série de divisions et de rapprochements. Cette fonction discursive est analogue à celle qu'occupait la *dianoia* dans le système de Platon. La deuxième raison pour laquelle la synthèse reste proche de l'intelligible, c'est qu'elle intègre des schèmes. Ces derniers ne sont ni des images ni des formes de raisonnement, mais une sorte de contemplation des concepts. Cette contemplation cependant n'est pas immédiate, comme l'était la *noësis* de Platon. Elle est un processus, une construction mentale des objets sensibles. En bref, Kant réinscrit l'intelligible dans le sensible, mais sans en effacer les nobles contours. On peut dire qu'avec lui le ciel des Idées nous est tombé sur la tête. L'imagination s'est sublimée ; l'intuition s'est faite chair.

2.5 La Raison : L'Homme qui pense qu'il pense

Tant que les catégories de l'Entendement s'exercent sur les intuitions sensibles, la connaissance reste fondée *sur* l'expérience *par* les principes *a priori* de l'esprit. Mais lorsque la pensée pure, indépendamment de tout contenu, s'élève aux idées transcendantales selon le fil déductif que la logique lui autorise, elle se perd bientôt dans des spéculations abstraites et antinomiques. Si l'Entendement cherche à ramener l'unité de la sensation à certaines règles, la raison pour sa part cherche à ramener l'unité de ces règles à certains principes.

tasma et *phantasia*. La première notion réfère à des *doxas* représentées indépendamment de toute sensation (des productions de l'imagination), tandis que la *phantasia* doit plutôt être entendue comme une déformation de la perception sensible. Pour plus de détails sur le sujet, voir notamment COLLETTE [?], *Phantasia et phantasma chez Platon*. Notons que certains auteurs (Baccou, Cazeaux, Karsenti/Prélorentzos) traduisent aussi "l'Eikasia" par imagination (voir section 1.2 *L'Analogie de la ligne*).

42. cf. KANT [?], *L'Anthropologie d'un point de vue pragmatique*. Le chapitre § xxviii est intitulé *De l'imagination* et le chapitre § xxx, *De l'imagination productive*.

43. Contrairement à Platon qui n'admettait dans l'art que ce qui participait à l'éducation morale et à la vérité.

Kant distingue trois idées transcendantales : *l'âme*, le *monde* et *Dieu*, étudiées respectivement par la *psychologie*, la *cosmologie* et la *théologie* rationnelles. Ces idées sont obtenues à partir des trois formes de raisonnements (ou plus précisément des séries prosyllogistiques de raisonnements) que sont le raisonnement *catégorique*, *hypothétique* et *disjonctif*. Kant analyse les sophismes découlant de chacune de ces trois espèces de raisonnements de façon à circonscrire les limites de validité de la raison. Nous n'allons pas entrer dans les détails de chacune de ces discussions, sans quoi cela nous éloignerait par trop de notre sujet. Nous allons seulement exemplifier la démarche dans le cas de l'âme qui, d'entre les trois antinomies, est la plus facile à mettre en relation avec la tradition socratique.

Selon la psychologie rationnelle, le concept de l'âme peut être obtenu en raisonnant à partir du « Je pense », sans qu'il soit besoin de se référer à l'expérience ou à une intuition sensible. Les principales propriétés de l'âme sont directement dérivées de la table des catégories, en appliquant les concepts purs de l'Entendement à l'*ego cogito*⁴⁴ :

	1	
	L'âme est	
	une substance	
2		3
Simple quant		Identique au
à sa qualité		cours du temps
	4	
	En rapport avec des objets	
	possibles dans l'espace	

Du premier de ces prédicats, la psychologie rationnelle tire l'idée de l'immatérialité de l'âme ; du second, celle de son incorruptibilité ; du troisième, celle de sa personnalité. Lorsque les trois sont pris conjointement, il vient de l'âme qu'elle est spirituelle et immortelle. Enfin, le quatrième prédicat donne l'idée de l'âme comme d'un principe vital, à rapprocher du concept de l'anima chez les Grecs.

Cependant, ces raisonnements ne sont, aux yeux de Kant, que des paralogismes. Le « Je pense » en effet ne me permet pas de me prendre pour objet de connaissance, et ce pour deux raisons : 1) d'une part il n'y a pas dans la pensée pure d'intuition sensible de moi-même, et par suite les idées qui en découlent ne sont que des formes vides ; 2) d'autre part je ne peux me connaître dans le « Je pense » qu'en tant que je suis l'unité de la perception, non en tant que substance. C'est donc par erreur que l'on s' imagine pouvoir déduire quoi que ce soit de l'être du moi à partir de l'*ego cogito* de la pensée pure.

44. KANT, *Critique de la raison pure*, trad. de BARNI, Dialectique transcendantale, Les paralogismes de la raison pure, p. 343

Deux remarques peuvent être tirées de cette analyse⁴⁵. Tout d'abord, l'on observe que la théorie kantienne n'autorise pas une philosophie de la subjectivité en tant que telle. Le sujet n'est pas à même de se prendre lui-même comme objet de connaissance, ou lorsqu'il le fait, il se présente comme une apparence (dans le cas de la sensibilité), comme une simple unité de conscience (dans le cas de l'Entendement) ou comme une illusion (dans le cas de la raison). En aucun cas l'on ne peut parler d'un sujet réflexif, capable de se donner son propre concept et d'entrer par ce fait dans une dynamique de dépassement de soi. Au contraire, le système kantien est baigné de statisme et de formalisme. L'antique γνῶσις σεαυτόν (*connais-toi toi-même*) prôné par Socrate ne peut donc plus s'appliquer. Dès lors, la démarche de la connaissance n'implique plus de devoir se transformer, en tant que sujet, dans son être même. Loin des rites d'initiation réalisés dans les couloirs des sectes pythagoriciennes et orphiques, loin de l'éducation aux idées et de la sortie de la caverne de Platon, le système kantien invite le sujet à faire entrer l'objet de son savoir dans un laboratoire. Il ne s'agit plus de changer au contact de cette nature qui nous fait face, et dont on tente de percer les mystères, mais de rendre l'objet de l'expérience identique à soi-même, de le soumettre à nos principes. Nous ne trouverons en lui rien d'autre que ce que l'Entendement y aura lui-même placé. Cette logique de l'identique amorce une sortie des spiritualités et l'entrée dans la connaissance rationnelle et dans la philosophie du Même.

Une autre remarque d'importance majeure, c'est que, contrairement à Platon, les structures innées de l'esprit humain ne sont plus définies sur base d'un modèle métaphysique. Pour Platon comme pour Kant, l'esprit possède des connaissances préalables à toute expérience, qui sont déjà en lui et qu'il ne fait que (re)découvrir. Chez Platon, cela est rendu possible du fait que l'âme, qui est immortelle, a déjà contemplé les idées, mais en a perdu le souvenir lors de son incarnation. Le savoir consiste donc en une réminiscence ; l'éducation est une maïeutique. Apprendre, c'est accoucher de ses oublis. C'est pourquoi chez Platon la Raison est une contemplation (*noësis*), en soi, de ce qui est toujours déjà hors de soi. Pour Kant au contraire, le savoir doit se dispenser de tout fondement métaphysique. La Raison pure ne conduit qu'à l'erreur et aux égarements. S'il existe bien des vérités *a priori*, c'est parce que le réel en entier apparaît conformément à nos dispositions intérieures, que sont nos intuitions pures et nos catégories. L'outil approprié n'est plus à chercher dans la contemplation ni dans le libre exercice de la pensée pure, mais dans l'Entendement. Cette faculté fonctionne à la manière de la *dianoia* de Platon, en raisonnant toujours à partir d'hypothèses et d'expériences possibles, mais avec le concours des intuitions sensibles. Pour Kant, il s'agit donc d'intuitionner hors de soi ce qui est toujours déjà

45. Ces deux remarques sont de moi, comme l'analyse et les résumés de Kant. Je me suis uniquement basé sur la lecture des sources originelles.

en soi.

2.6 La Beauté sans vérité

Quelle est la nouvelle conception du jugement qui se dégage de la révolution kantienne ? On se souviendra que, chez Platon, le Vrai, le Beau et le Bon étaient les trois visages divins, les trois facettes complémentaires d'une même essence fondamentale — que Socrate comparait au Soleil. Contempler une de ces Idées pleinement et authentiquement, c'est jeter un coup d'œil sous la surface d'un lac. Tant qu'on garde la tête hors de l'eau, on croit voir des rochers isolés, des souches avalées par la vase, des algues solitaires. Mais lorsqu'on plonge la tête dans l'eau, on se rend compte que ces rochers, ces souches, ces herbes forment un seul et même tout. On ne peut pas poser les yeux sur l'un de ces éléments sans embrasser du regard tout ce qui les relie. Chez Platon, la vérité, le bien moral et la beauté sont liés si étroitement, si vivement, qu'il est impossible de faire une action belle qui ne soit pas aussi vraie et bonne. N'est-ce pas d'ailleurs la raison pour laquelle les anciens Athéniens disaient, avec force conviction, que la vie bonne est aussi, et avant tout, une vie belle ?

C'est cette solidarité immuable des Idées à laquelle Kant s'attaque au travers de ses diverses critiques de la raison pure, de la raison pratique et de la faculté de juger, et ce de trois manières. Tout d'abord, (1) il retire à la Raison spéculative le privilège de la vérité ; de plus, (2) il lègue à la Raison le pouvoir de déterminer le bien du mal ; et enfin, (3) il voit le beau comme un libre exercice de l'imagination sensible. Le premier de ces trois points est purement *critique* et *limitatif*. Kant, on s'en souvient, montre que les grandes questions concernant l'existence de Dieu, la finitude de l'Univers ou l'immortalité de l'âme, sont indécidables. Les concepts mêmes qui sont mobilisés dans ces questions sont insuffisants, si bien qu'il est possible de prouver une chose et son contraire. La vérité de la science, pour Kant, ne dépend pas purement du jugement analytique, mais de l'expérience du monde. Cette expérience est cependant modelée par nos concepts, par les structures innées de notre esprit, si bien que les vérités scientifiques sont avant tout des régularités de la perception humaine.

En refusant à la Raison sa capacité à penser le vrai, en fondant la vérité des sciences sur les formes *a priori* de la Sensibilité et de l'Entendement, la critique kantienne bouleverse l'échafaudage de la métaphysique. Dorénavant, il est inutile de prétendre à cette contemplation, à cette *noesis* mystique, presque ésotérique, qui était au fondement de l'école de Platon. Le but de la philosophie, désormais, n'est plus de se transformer de l'intérieur, ni d'élever son âme vers le lieu intelligible. Dorénavant, il s'agit de critiquer, au moyen de la méthode transcendantale, les illusions de la métaphysique.

Bien que Kant nie que la Raison puisse nous enseigner quoi que ce soit

à propos des questions métaphysiques spéculatives, il ne lui enlève pas le privilège de juger du bien et du mal. Il s'agit là du point (2). Dans la *Critique de la raison pratique* et dans la *Fondements de la métaphysique des mœurs*, Kant entreprend d'élucider les lois *a priori* de la morale. Ces dernières sont fondées sur le principe que l'être humain est une fin en soi. Si la nature a mis en nous cette étrange capacité à raisonner, alors qu'elle aurait eu beau jeu de nous orienter avec le bâton et la carotte des plaisirs et des peines, c'est que la raison doit bien servir à quelque chose. Il faut respecter, en chacun, ce miracle proprement extraordinaire qui est celui de la pensée.

C'est cette capacité de penser qui est au fondement de notre dignité humaine, c'est tout le sel et le piquant de la vie, la corroboration d'une nature humaine décidément mystérieuse. Tant qu'un être est capable de penser logiquement, et universellement, il ne peut pas nier que toute autre personne capable d'en faire autant mérite, tout autant que lui, d'être respectée comme une fin en soi. L'exercice de la morale est donc rendu possible par l'adéquation de nos agissements à cet impératif fondamental^[46] :

« Agis de telle sorte que la maxime de ton action puisse être érigée par ta volonté en une loi universelle ; agis de telle sorte que tu traites toujours l'humanité en toi-même et en autrui comme une fin et jamais comme un moyen ; agis comme si tu étais à la fois législateur et sujet dans la république des volontés libres et raisonnables. »

La morale se fonde donc, non pas sur la compassion ou sur l'intérêt social, mais sur le devoir désintéressé. Ce devoir consiste à respecter l'autonomie de la raison et, au travers d'elle, de tous les êtres raisonnables. On peut donc dire que le bien, chez Kant, est déontologie. Il s'appuie sur des lois que le raisonnement doit nous permettre d'élucider. Par ailleurs, ce raisonnement n'est pas de l'ordre d'une contemplation ou d'une transformation intérieure. Il relève davantage d'une discipline, d'une droite méthode. Il y a quelque chose d'austère, de désincarné, dans cette vision du bien, qui n'aurait pas trouvé écho chez les anciens Athéniens pétris des mystères d'Éleusis, des bacchanales, des rites pythagoriciens. Est-ce à dire que Kant se désintéresse totalement du Beau ? Qu'il cesse d'y voir un lien privilégié, une porte d'accès, un apprentissage devant mener à la vérité supérieure ?

Dans plusieurs de ses écrits, Kant défend au contraire que, bien que la morale tire son origine dans la raison seule, le sentiment moral, quant à lui, participe de sa force et de son caractère sublime. L'émotion ressentie devant l'héroïsme, tout particulièrement l'héroïsme sacrificiel, ou la vertu exercée sans considération pour l'intérêt personnel, est à ses yeux une constante de la nature humaine^[47]. Loin d'y voir une exaltation purement germanique, où se

46. La citation ci-dessous est une contraction des trois impératifs kantien, tels qu'ils sont formulés dans la *Fondements de la métaphysique des mœurs*.

47. Voir la *Fondements de la métaphysique des mœurs*, où Kant écrit : "Car l'observation

mêlent des notions de patriotisme, il défend au contraire qu'il s'agit là d'une émotion universelle, le socle constitutif, le terreau fertile du développement moral précoce. Il faut donc entendre que le Beau, bien qu'il soit séparé du Bien, n'est pas pour autant abandonné. Il continue de jouer un rôle central dans le jugement humain.

Ce rôle, et il s'agit du point (3), n'est toutefois pas celui que lui donnait Platon. Pour ce dernier, la beauté était une sorte de fil d'Ariane, abandonné dans le labyrinthe de la vie à l'adresse des plus sages, pour les guider peu à peu sur le chemin de la vérité et du bien supérieur. C'est ainsi que les jeunes gens, tout éblouis qu'ils sont encore par la vivacité des sens et du plaisir, poursuivent tout d'abord les impulsions de l'Éros. Ils apprennent à aimer les beaux corps, à les embrasser, à les posséder, avant de se tourner vers cette beauté qui n'a pas de forme, pas de visage, pas de matière. Ils s'élèvent ainsi, étape par étape, vers la contemplation intellectuelle, en commençant par l'illusion et la croyance, avant d'arriver au plein exercice de la raison.

Chez Kant, au contraire, le beau n'est pas une fin en soi. Il consacre à la question du beau son troisième ouvrage majeur, la *Critique du jugement*. Il y défend l'idée que l'expérience esthétique est le jeu désintéressé de l'imagination et de la sensibilité. Par *désintéressé*, il faut entendre que ce qui est beau n'a pas d'utilité et n'est pas agréable. Ce qui naît d'une passion amoureuse ou d'un intérêt charnel n'est beau que dans la mesure où il procure du plaisir. Il se peut d'ailleurs tout aussi bien que ce corps que l'on jugeait beau, dès lors qu'il nous refuse ses faveurs, nous paraisse tout à coup haïssable. Cette transition, ce chavirement bien connu des poètes, qui transforme l'adoration en haine et la haine en passion, est le signe que ce qui est en jeu, c'est nous-mêmes : notre intérêt, notre plaisir, notre futur — et non pas la beauté en tant que telle.

En d'autres termes, Kant pose d'emblée que la beauté n'est pas une fin et ne saurait, à ce titre, avoir quoi que ce soit en commun avec la raison. C'est d'ailleurs pour cela qu'il est difficile, voire impossible, d'exprimer en des termes raisonnables pourquoi une chose est belle plutôt que d'être laide. Lorsqu'on tente de le faire, on perd de vue l'essentiel de la beauté. On trouve des explications, on dit que les formes sont agréables, que les couleurs sont douces, que les mélodies sont délicates ou que l'artiste à l'origine de ces œuvres esthétiques possède telle et telle qualité.

Il ne faut toutefois pas tomber de Charybde en Sylla : bien que le ju-

la plus commune montre que si l'on présente un acte de probité détaché de toute vue d'intérêt quel qu'il soit, en ce monde ou dans l'autre, accompli d'une âme ferme même au milieu des plus grandes tentations que fait naître le besoin ou la séduction de certains avantages, il laisse bien loin derrière lui et éclipse tout acte analogue qui dans la plus petite mesure seulement aurait été affecté par un mobile étranger, qu'il lui élève l'âme et qu'il excite le désir d'en pouvoir faire autant. Même des enfants d'âge moyen ressentent cette impression, et l'on ne devrait jamais non plus leur présenter les devoirs autrement." (p.26)

gement de goût soit non rationnel, il a quelque chose d'universalisable. Je dis bien universalisable, et non universel, car les goûts et les couleurs, selon l'adage bien connu, ne sauraient être les mêmes pour toutes et tous. En revanche, celui qui pose un jugement favorable, en décrétant de telle chose qu'elle est belle, pense invariablement que la beauté est une propriété objective de la chose. Bien que le sentiment du beau vienne de lui, il en parle comme si toutes et tous *devaient* nécessairement éprouver la même chose. Il y a donc une propension, dans l'expérience esthétique, à la généralisation — propension qui ne se fonde pas, contrairement à la morale, sur la nature universelle de la raison.

Quel est donc, dès lors, cet ingrédient, propre au jugement de goût, qui tend à la généralisation ? Kant en identifie deux : l'imagination et les catégories de l'entendement. Ces aptitudes ne sont toutefois pas mobilisées de la même façon que lorsqu'il s'agit de former des concepts. Dans ce dernier cas, en effet, on a vu que la catégorie permet de donner aux phénomènes une structure logique, en séparant les objets les uns des autres, en leur appliquant la loi de causalité, en distinguant le possible de la nécessité. L'imagination, de même, joue un rôle clef dans la schématisation. Elle permet de se représenter un concept abstrait — un triangle — au moyen d'une forme confuse, vaguement colorée et stéréotypée, qui ressort par contraste sur le blanc d'une page vierge.

Or, dans le cas du jugement de goût, il n'y a pas de concepts. Kant est formel sur ce point. Ce n'est pas la représentation de la pomme au milieu du jardin d'Éden, qu'une femme alanguie tient entre ses mains tendues, qui doit être qualifiée de belle. Ce qui est beau, ce sont justement les portraits artistiques qui déforment les concepts, qui les idéalisent, qui se jouent des conventions ordinaires. Il y a un processus profondément créateur, dans l'esprit humain, qui atteint son apogée dans l'expérience esthétique. Loin de se tenir à la façon d'un réceptacle passif, se contentant de recevoir, en lui, la marque de la beauté extérieure ; le sujet ébahi par la beauté participe activement à la sensation d'émerveillement.

À ce premier type d'expérience esthétique, que l'on pourrait qualifier de libre et de créatif, se tient également l'expérience du sublime. Cette dernière se rencontre, encore une fois, lorsque la chose représentée ne correspond pas à son concept, mais dans un cas tout particulier : lorsque ce qui est exprimé est trop grand, trop vif, trop immense, pour être circonscrit dans une représentation ; lorsqu'un paysage, une tempête, un sentiment d'héroïsme, semble défier les frontières bien sages de nos concepts, alors seulement l'esprit est saisi de son infirmité. Il succombe à l'immensité du monde. Il tremble, il palpite, il résonne à l'unisson avec ce tout qui le dépasse. Au cœur du beau, en ce qu'il a de plus noble et de plus renversant, se tient ainsi l'idée de l'infini.

Pour résumer la conception du beau, telle qu'elle se dégage à l'issue de la critique kantienne, on pourrait dire qu'elle est déliée des entraves qui la retenaient auprès du bien et du vrai. Aux trois facultés kantienne, que sont la

Sensibilité, l'Entendement et la Raison, échoient l'une des trois idées platoniciennes. Le Bien devient une loi purement rationnelle, logique, universelle. Le vrai est donné par l'expérience sensible et informée par les catégories de l'Entendement. Le Beau quant à lui, est cet au-delà de la représentation, cette autonomie du monde sensible, cette hébétude de l'infini, qui défie les limites de l'Entendement.

Dans la deuxième partie de cette thèse, je tenterai de montrer comment la pensée moderne, incarnée par le fonctionnement des machines, prolonge cette déconstruction de la tripartite platonicienne. Avec l'avènement du calcul et des formalismes logiques, des langages binaires qui sont au fondement de l'informatique, la rationalité devient le seul critère du vrai. Ce qui est valide doit faire l'objet d'une déduction logique, automatique. On se méfie des observations, des intuitions, des instincts qui n'ont pas préalablement été informées par un cadre méthodologique, scientifique et calculatoire.

Ce mouvement est ce que j'appellerai l'oubli de l'Entendement. Il consiste à mettre à plat la complexité de l'esprit humain — et du monde qui se donne au travers de lui. On ne voit plus désormais que deux niveaux de pensées, deux facultés : la sensation et la rationalité, les jugements synthétiques *a posteriori* et les jugements analytiques. Le jugement de goût devient une affaire personnelle, sans vocation à l'universalité. La vérité se réduit à des jeux de corrélations. Le bien n'est plus qu'une règle abstraite, sans fondement dans le cœur de l'homme. Il n'y a plus de liaison, même partielle, entre le beau, le bien, le vrai.

Cet oubli de l'Entendement, cet abandon de la beauté, ce relativisme de la vérité, explique en partie la confiance sacro-sainte que nous accordons aux algorithmes. Ces derniers sont le produit d'un tournant historique, d'une épuration idéologique, entamée dès l'entrée dans la modernité, qui consiste à dépouiller tout ce qu'il restait de transcendant et de contemplatif dans l'expérience de l'Homme.

La deuxième partie, intitulée *Révision du Concept de jugement*, consistera donc à expliquer comment on en est arrivé à oublier la beauté et en quoi cet oubli affecte le jugement des machines que nous créons. Avant toutefois d'entamer le procès méthodique de la rationalité algorithmique moderne, j'aimerais utiliser les outils philosophiques développés jusqu'à présent, au travers de Kant et de Platon, pour faire le point sur notre trajectoire philosophique : au vu de ce qui a été établi jusqu'à présent, peut-on dire qu'une machine ait la capacité de juger ?

Le prochain chapitre distinguera plusieurs types de jugements. Je tenterai de déterminer si les algorithmes peuvent faire des jugements rationnels ? Des jugements basés sur l'expérience ? Des jugements synthétiques *a priori* ? Ces ébauches de réflexions seront l'occasion de se demander si une machine est dotée d'une intuition du temps et de l'espace, si elle possède des concepts, de l'imagination et un certain sens subjectif d'être un sujet, un point de vue en Je posé sur le monde. Elles serviront de point de départ aux réflexions

qui suivront. Mais auparavant, il me faut encore faire quelques remarques méthodologiques concernant l'importance que j'ai consacrée, tout au long de ma thèse, à la première critique kantienne.

2.6.1 L'Importance de la première critique

Un lecteur consciencieux pourrait me demander pourquoi j'ai pris le parti, dans cette thèse, de me concentrer sur la Critique de la Raison pure, sans donner à la deuxième critique (celle de la raison pratique) et à la troisième critique (celle du jugement esthétique), une place plus importante.

Outre ce zeste d'arbitraire, subtilement mêlé aux préférences subjectives, qui plane comme un spectre invisible sur le choix d'une littérature de référence, je voudrais montrer quelques éléments objectifs qui faisaient de la première critique la candidate idéale pour mon approche.

À cette fin, peut-être n'est-il pas inutile de rappeler l'objectif stratégique de cette analyse de la pensée kantienne ? Loin de faire une simple archéologie descriptive de sa philosophie, à la façon d'un érudit ou d'un scolastique pétri du désir de montrer ce qu'il sait, mon point central consiste en ceci : montrer qu'il existe une continuité, une paternité pourrait-on dire, entre le concept du jugement, tel qu'on le trouve chez Platon, et le concept de jugement chez Kant.

Plus précisément encore, je voulais montrer que cette pensée à deux niveaux qu'on trouve chez Platon, qui oppose la sensation et la raison, peut acquérir une épaisseur supplémentaire en glissant, entre ces deux facultés, celle de l'Entendement. C'est exactement le mouvement inaugural orchestré par la Critique de la raison pure de Kant, et ce mouvement, du reste, est réalisé dans la pleine conscience du crime.

Au moment de discuter les concepts de l'Entendement, j'ai montré que Kant fait une référence explicite à « la Colombe qui traverse l'air sans rencontrer de résistances⁴⁸ » (voir la section intitulée « La Métaphore du Colombier »). Ce qu'il désigne par cette phrase lapidaire, c'est la métaphore du Colombier de Platon, où le jugement est dit vrai ou faux de façon idéale, sans faire référence au contenu de la sensibilité.

Il était crucial pour Kant, au moment d'introduire le jugement de l'entendement, de montrer qu'il existait quelque chose entre la raison pure envisagée par Platon — consacrée à l'étude de l'inconditionnel — et la sensation proprement dite. Si cette référence à Platon vous semble incertaine, considérez encore ceci. Kant reprend plusieurs fois l'exemple que « $5 + 7 = 12$ » (Théétète, 196a). Cet exemple ouvre la seconde préface de la Raison pure pour illustrer le jugement synthétique a priori. Or il se trouve qu'il nous vient, une fois encore, du passage du Théétète parlant du Colombier.

Cela devrait, je crois, justifier pourquoi j'ai choisi de me concentrer sur

48. *Ibidem*, Introduction III, pp. 52-53.

la *Critique de la raison pure*, et non sur les autres ouvrages de Kant. Il ne s'agit bien entendu pas de la seule manière de fonder une critique de l'intelligence artificielle — loin s'en faut ! La dichotomie opposant le jugement déterminant et le jugement réfléchissant, que Kant développe dans sa *Critique de la faculté de juger*, pourrait s'avérer intéressante. Je pense néanmoins que ces concepts ont moins de pertinence que le système à trois niveaux de l'entendement, de la sensibilité et de la raison.

Le jugement déterminant désigne en effet, en langage moderne, l'approche déductive, où un jugement est construit à partir d'un principe universel. Dans le syllogisme « Tous les hommes sont mortels et Socrate est un homme ; donc Socrate est mortel » , on passe d'une vérité universelle (tous les hommes sont mortels) à une vérité particulière (Socrate est mortel).

Le jugement réfléchissant, quant à lui, se rapporte plutôt à une induction. Loin de partir des catégories de la raison, il se concentre sur les phénomènes. Il cherche à les étudier, à les faire parler, pour dégager des vérités plus générales. Il fonctionne pour ainsi dire à l'aveugle, sans déterminations préalables.

À première vue, ces distinctions pourraient correspondre, quoique de façon certes grossière, à la distinction que j'ai introduite, dans l'introduction de cette thèse, entre l'approche top-down et bottom-up. Les stratégies top-down, celles qui étaient le plus en vigueur aux premiers temps de l'intelligence artificielle, consistent à définir un ensemble de principes généraux et, à partir de ceux-ci, de s'acheminer à des conclusions. L'exemple archétypal est le fameux système expert, dont je toucherai un mot dans la section [6.3](#).

Les stratégies bottom-up, en revanche, partent des données de l'expérience et affinent leurs décisions à leur contact. Elles relèvent d'un apprentissage sans supervision humaine, ou avec une supervision relativement contrôlée. C'est l'ambition des programmes d'intelligence artificielle de la deuxième génération : celle du *Deep Learning* et des réseaux de neurones — dont ChatGPT est l'aboutissement le plus spectaculaire.

On pourrait critiquer ce parallélisme, certes rapide, que je tisse entre l'approche bottom-up et le jugement réfléchissant, en se fondant notamment sur les concepts de liberté ou de jugement esthétique, qui sont indispensables à Kant — et qui paraissent tout à fait subordonnés aux processus de calcul dans le cadre des algorithmes actuels. Il n'en reste pas moins que cette opposition entre jugement réfléchissant et jugement déterminant ne présente que deux niveaux, et non trois. Elle ne sert donc pas ma démarche de façon tout à fait prometteuse.

Ces remarques devraient, je crois, achever de montrer que la construction de cette thèse, si elle n'échappe pas tout à fait à l'arbitraire, suit une logique intrinsèque que n'aurait pas autorisé — toutefois pas si aisément — une carte de lecture focalisée sur la deuxième ou la troisième critique kantienne.

Kant critique les conceptions de Platon et élabore une théorie du jugement fondé sur trois facultés :

1. *La Sensibilité* est l'aptitude à réaliser des jugements d'expérience ; cette synthèse s'appuie néanmoins sur des formes préalables de la pensée humaine que sont l'intuition du temps et l'intuition de l'espace.
2. *La Raison* est l'aptitude à réaliser des jugements rationnels. Cette analyse n'est fondée sur aucune donnée d'expérience ; elle se contente de penser les articulations purement logiques d'une idée.
3. *L'Entendement*, quant à lui, est l'aptitude à réaliser des jugements synthétiques *a priori*. Ce jugement un peu particulier est une révision du jugement sans justification de Platon. D'une part, il repose sur les structures même de l'expérience (les intuitions du temps et de l'espace) mais, d'autre part, il touche à des concepts universels (les douze catégories de l'Entendement).

On comprend donc que l'Entendement est un liant essentiel entre la synthèse et l'analyse ; entre l'expérience et la logique. J'ai suggéré en outre que l'Entendement puisse être l'ingrédient faisant défaut à l'ordinateur. Face aux exercices ARC de François Chollet, les algorithmes peinent en effet à généraliser et à abstraire leur observations.

Dans le prochain chapitre, je vais discuter plus en détails les différences de structures séparant l'être humain et nos algorithmes actuels au sein d'un cadre d'analyse kantien.

Chapitre 3

Le Jugement artificiel : La Machine peut-elle juger ?

*Nous sommes comme cet homme
qui est monté sur une chaise et
a déclaré qu'il était un peu plus
proche de la lune.*

Hubert Dreyfus

Il est maintenant temps de se demander ce que la théorie kantienne peut nous enseigner à propos de l'intelligence artificielle. Au début de notre discussion sur Kant se trouvait en effet une pierre d'achoppement, que nous avons tenté de contourner adroitement. Cette pierre, ce sont les tests ARC de François Chollet, face auxquels les algorithmes actuels présentent des résultats plus que médiocres (voir figure 3.1).

Ces exercices sont similaires aux matrices de Raven, mais moins standardisés et plus créatifs. Ils consistent à généraliser certaines situations sur la base de deux ou trois exemples seulement. On a vu qu'il était aisé pour l'être humain de donner la bonne réponse, mais pas pour les algorithmes. Tout se passe comme si, pour nous autres, les schémas de généralisation étaient inscrits dans nos structures perceptives. Ces structures ne semblent pas dépendre beaucoup de l'expérience ; bien qu'elles se développent au contact du monde, elles sont présentes en nous de façon innée. Elles nous permettent de saisir les symétries spatiales, les rotations et les opérations de pensée abstraites figurant dans les théorèmes de la géométrie euclidienne, lorsque l'on dit, par exemple, que la somme des angles fait 180 degrés.

Le problème qui se présente dès lors est de savoir si la pensée abstraite requise pour résoudre les problèmes ARC ne s'apparenterait pas aux jugements synthétiques *a priori* que j'ai introduits au précédent chapitre ? Dans l'affirmative, nous détiendrions un éclairage intéressant sur ce qui différencie le jugement de l'être humain et de l'intelligence artificielle. Le jugement

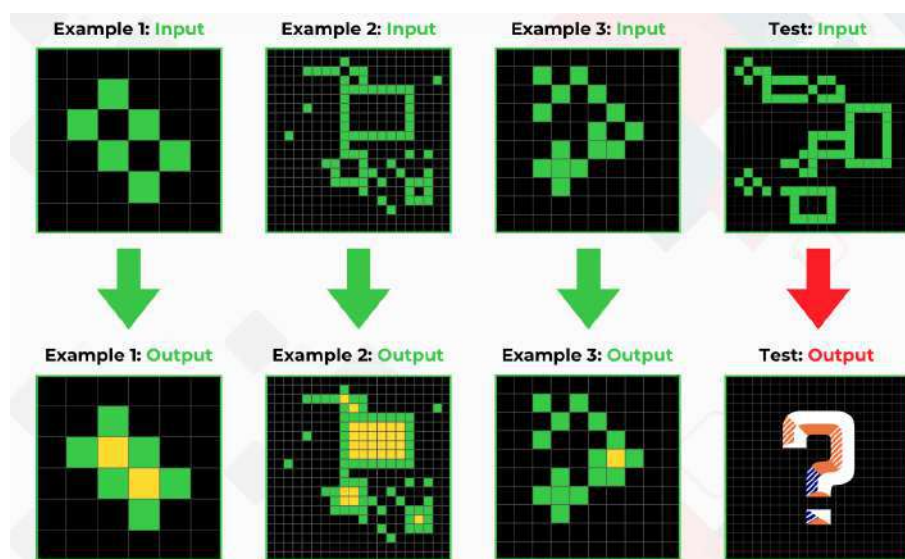


FIGURE 3.1 – Voici un exemple de test ARC. À partir d'un input donné, la machine doit fournir le bon output. Elle reçoit trois exemples pour comprendre ce qu'elle doit faire (source : <https://lab42.global/arc/>).

synthétique *a priori*, en effet, repose en substance sur des formes pures. Si l'on en croit la théorie de Kant, elles existeraient en nous préalablement à toute expérience : ce sont les intuitions de l'espace et du temps (chapitre 2.3), ainsi que les catégories de l'Entendement (chapitre 2.4).

Contrairement aux capteurs d'un robot, qui se contentent de saisir les données de l'environnement de façon passive, ces formes *a priori* ne sont pas de simples réceptacles. Ce sont des structures, des filtres, qui *informent* le monde, qui le *déforment* à l'occasion. Autrement dit, nous sommes équipés de concepts géométriques qui déterminent notre façon de percevoir le monde, nous permettant d'isoler des objets, d'abstraire des propriétés et de poser des conclusions logiques même en présence de problèmes sur lesquels nous n'avons pas été spécifiquement entraînés (ou fort peu).

La question que j'aimerais soulever dans la suite de ma thèse consiste à déterminer si, au sein de l'esprit humain, ce ne seraient pas les données sensibles qui s'adaptent à nos concepts *a priori*; tandis qu'au sein des algorithmes, ce seraient les concepts qui cherchent, péniblement, à s'adapter aux données? Cette idée que j'évoque expliquerait que nous saisissons les matrices de Raven plus facilement. Ces matrices sont en effet conçues pour répondre à nos structures de pensée. Lorsqu'on nous les présente, nous nous attendons à voir des symétries, des translations, des carrés, des rectangles. La machine, au contraire, qui n'a pas d'*a priori*, doit tenter de saisir ces transformations en se basant uniquement sur des exemples. Comme ceux-ci sont en nombre insuffisant, elle produit des réponses erronées.



FIGURE 3.2 – Dans les illusions de Kanizsa, une figure géométrique apparaît dans le blanc de la page. D’un point de vue kantien, cela illustre que c’est le monde sensible qui s’adapte à nos catégories, et non l’inverse.

Pour illustrer la présence, chez l’être humain, de concepts géométriques structurels, j’aimerais me pencher sur certaines illusions d’optique, notamment l’illusion de Kanizsa (voir figure 3.2). Dans cette dernière, des disques noirs sont creusés de sorte à suggérer, dans le blanc de la page, la présence d’un carré ou d’un triangle. Ces formes que nous avons l’impression de voir sont un pur produit de l’imagination. Tout se passe comme si notre esprit dessinait les contours d’un objet inexistant. Ces arêtes imaginaires que nous observons font écho à la théorie du schématisme kantien. Dans cette dernière, rappelons-le, l’esprit humain n’est pas un simple réceptacle passif. Il exerce un rôle actif dans la perception. L’imagination est mobilisée pour découper le réel et construire les contours des concepts auxquels nous sommes les plus familiers. C’est ce qui fait dire à Kant que « Le schème n’est toujours par lui-même qu’un produit de l’imagination [...] servant à procurer à un concept son image¹. »

Lorsqu’on présente l’illusion de Kanizsa à un logiciel tel que ChatGPT, les résultats sont mitigés. L’algorithme est certes capable d’identifier la présence du triangle et du carré dans les situations initiales, mais si on lui présente une configuration nouvelle, en effectuant une perturbation aléatoire du schéma d’origine — et qui n’existe pas sur Internet — on obtient une réponse fort différente. La discussion ci-dessous a été obtenue en soumettant l’image ci-contre (figure 3.3) à ChatGPT² (version o1 de novembre 2024) :

1. KANT, *Critique de la raison pure*, Tome I, Du schématisme des concepts purs de l’entendement, chapitre 1, p. 201.

2. Cette expérience a été réalisée par Patrick Bestgen, Ingénieur en Informatique et Directeur de Recherche, Consult’It Suisse. Il a lu les brouillons de ma thèse et s’est proposé de tester directement mon idée avec ChatGPT.

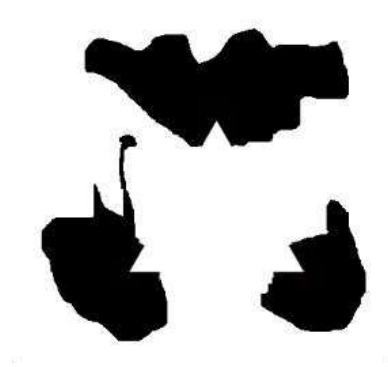


FIGURE 3.3 – Illusion de Kanizsa après avoir effectué une modification chaotique aléatoire. Nous voyons toujours le triangle, mais pas ChatGPT.

HUMAIN : Une figure géométrique apparaît dans le blanc de la page. Quelle est-elle ?

ChatGPT : La figure géométrique qui apparaît dans le blanc de la page est une étoile à six branches. Elle est formée par l'espace négatif entre les formes noires plus amorphes.

HUMAIN : Moi, je ne vois pas d'étoile à six branches. En tout cas, elles ne sont pas symétriques ! Ne vois-tu pas une autre forme plus évidente ?

ChatGPT : Tu as raison, si l'on cherche une symétrie parfaite, l'étoile à six branches n'est pas la forme la plus évidente. En regardant l'espace blanc, je vois plus clairement un triangle avec ses pointes arrondies/incomplètes.

Il est certes permis de croire que ChatGPT, avec un entraînement suffisant, finira par apporter des réponses adéquates. L'illusion de Kanizsa suggère néanmoins qu'il y a une différence de taille dans la façon dont les données apparaissent à l'être humain et dans la façon dont elles sont traitées par la machine. Bien que ce soient les mêmes ensembles d'objets, à savoir des points sur un écran, ils ne seraient pas vus au travers des mêmes filtres.

L'intelligence artificielle voit des points isolés les uns des autres ; et c'est toute une affaire de les relier, de les classer, d'identifier dans ces nuages disparates la forme d'une lettre, d'un visage ou d'une figure. Tout au plus peut-elle procéder à des associations statistiques aveugles, qui imitent les réponses attendues. L'esprit humain, quant à lui, voit des schémas géométriques généralisables ; il déforme même parfois sa perception pour la faire correspondre à la structure de ses concepts, en ajoutant par exemple

des lignes qui n'existent pas. Qui plus est, il ne peut pas ne pas voir ces formes géométriques apparaître dans le blanc de la page, même avec un effort conscient. L'illusion s'impose à lui de toute nécessité.

Cette position que j'envisage, comme quoi nos concepts *a priori* modifient notre perception, est au cœur de la pensée kantienne. Elle a d'ailleurs souvent été qualifiée de révolution copernicienne³. Elle occupe une place intermédiaire entre l'idéalisme, qui fait de la conscience l'étalon de la réalité, et l'empirisme aveugle, qui ne tient pour réel que l'expérience des sens⁴:

Il y a des choses qui nous sont données, en tant qu'objets de nos sens situés hors de nous, mais, de ce qu'elles peuvent bien être en soi, nous ne savons rien, nous ne connaissons que leurs phénomènes, c'est-à-dire les représentations qu'elles produisent en nous en affectant nos sens.

Plutôt que de penser, somme toute naïvement, que nous percevons le monde tel qu'il est ; plutôt que de croire que l'Univers est peuplé de carrés, de triangles, de cercles, de matière, d'odeurs ou que sais-je encore, la position de Kant consiste à dire que nous ne voyons que ce qui correspond aux concepts qui sont en nous. C'est parce que nous avons, dans notre esprit, le concept du triangle, que des triangles nous apparaissent dans le monde tangible. Plutôt que de dire « Regarde, un triangle », nous ferions peut-être mieux de nous exclamer, à l'instar des peintes surréalistes : ceci n'est pas une pipe ? Existe-t-il en effet un seul triangle dans l'Univers ? Ce que nous prenons pour tel n'est-il pas, en dernière analyse, qu'un ensemble d'atomes en suspension dans le vide béant d'une page blanche ?

L'évolution de la science moderne semble apporter chaque jour la preuve que nos perceptions et nos concepts sont fort éloignés de la réalité physique. À l'époque de Kant, l'Univers était encore celui des poids et des balances, des atomes en collision, mus par les forces de Newton dans un espace euclidien, mais dès le début du vingtième siècle, nos idées préconçues à propos du temps, de l'espace, de la lumière, de la gravitation, éclatent tant et si bien que force est de constater que le monde d'aujourd'hui, la berge délicate de nos pensées, n'est plus fait que de rocs, de saillies, de trous.

Non seulement l'Univers est constitué presque intégralement de vide, mais les particules qui le peuplent sont victimes de dédoublement : leur compor-

3. C'est Kant lui-même qui introduit cette idée, bien qu'il ne cite pas le nom de Copernic. Ce que Kant a plutôt en vue, c'est la méthode d'Euclide en mathématique et celle de Galilée en physique. Il voit dans les démonstrations des *Éléments* d'Euclide et dans les plans inclinés de Galilée des espaces où « la raison n'aperçoit que ce qu'elle produit elle-même d'après ses propres plans » (*Critique de la raison pure*, préface de la seconde édition). Il ajoute plus loin : « C'est dans cette tentative de changer la méthode suivie en métaphysique et d'y opérer ainsi, suivant l'exemple des géomètres et des physiciens, une révolution complète, que consiste l'œuvre de notre critique de la raison pure spéculative » (*ibidem*).

4. KANT EMMANUEL, *Prolégomènes à toute métaphysique future*, §13, Remarque II.

tement, tantôt ondulatoire, tantôt particulière, implique une incertitude sur leurs trajectoires (vitesse ou position), si bien qu'on préfère parler de probabilité de présence ou d'orbitale quantique. L'espace-temps lui-même se complexifie : il se tord sous la présence de masses, se rétracte, se forme autour de trous noirs dantesques. Nous continuons pourtant de voir autour de nous des objets solides, dotés d'une couleur, d'une texture. Il faut donc que notre espèce soit dotée d'une certaine représentation intuitive du monde qui est bien différente de ce que la physique nous en donne à comprendre. Notre vue, cette éternelle joueuse, nous trompe sur la nature des choses, parce que nous sommes tout simplement incapables de voir les choses telles qu'elles sont.

Comment voyons-nous le monde ? Et en quoi exactement différons-nous des machines que nous programmons ? J'ai montré, dans les chapitres précédents, qu'il existait trois types d'*a priori* dans la pensée kantienne, qui progressaient main dans la main : (1) les intuitions, (2) les catégories de l'Entendement et (3) le Je.

Les intuitions relèvent de la Sensibilité. Ce sont les formes pures de l'espace et du temps, définies comme ce sans quoi il ne serait pas possible de se représenter un phénomène dans le temps et dans l'espace. Ce sont des préalables à toute expérience concrète du monde.

Les douze catégories de l'Entendement, comme leur nom l'indique, dépendent de l'Entendement. Elles permettent d'isoler au sein des phénomènes sensibles des objets dotés de relations, et mis en relations les uns avec les autres. Kant énumère quatre espèces générales de catégories : la quantité, la qualité, les relations et la modalité.

Le Je, pour sa part, est le fondement subjectif, le liant, le socle de toutes les perceptions. C'est ce sans quoi nos expériences de conscience ne seraient pas *nos* expériences à nous, ni même des *expériences* à proprement parler, mais de simples données isolées dans le temps et dans l'espace. Le Je est ce qui donne à notre conscience une unité subjective : tout ce que nous sentons, tout ce que nous ressentons, se rapporte à nous. Cette table, ce livre, ce fauteuil, le sentiment du vent sur notre joue, les odeurs de la pièce, les bruits de la rue, tout cela forme un seul et même tout.

Si c'est bien le jugement synthétique *a priori* qui est en cause, lorsqu'il s'agit de comprendre la différence entre le raisonnement humain et le raisonnement artificiel, il s'ensuit la chose suivante : l'IA échoue encore et toujours aux tâches ARC parce qu'elle ne détient pas les mêmes structures perceptives *a priori*. Elle analyse les données, mais ces dernières ne se manifestent pas, en elles, en suivant les mêmes structures. En quoi exactement peuvent consister ces structures ? Comment prouver que la machine ne les possède pas ?

Dans les pages qui suivent, je vais tenter de questionner, méthodiquement, dans quelle mesure une « machine » dispose — ou non — des structures *a priori* que nous avons identifiées. J'utiliserai le terme de « machine »

pour désigner, de façon générale, les outils de Deep Learning utilisés aujourd'hui, en ce compris les réseaux de neurones. Cette dénomination commune — vague, je vous l'accorde — ne devrait pas faire oublier au lecteur qu'il existe une grande diversité de méthodes algorithmiques, dotées chacune de leurs particularités. Une revue de la littérature de l'informatique théorique sera exposée dans les chapitres 7 et 6, au cours desquels je tenterai de développer une réponse rigoureuse aux questions que je ne fais, ici, que soulever.

Notons encore, avant de finir, que l'on pourrait critiquer la notion de jugement *a priori* donnée par Kant. Cette définition suppose en effet la présence de structures innées dont on peut se demander si elles existent. Je montrerai dans le chapitre 4 que les intentions de l'espace et du temps sont tout aussi problématiques lorsqu'on les envisage chez l'être humain que chez la machine. Il s'agira alors de construire une nouvelle définition du jugement qui, tout en tenant compte de l'apport critique de Kant, soit plus cohérente avec le développement récent de la science.

Pour le dire en court, ce que j'entreprends dans ce chapitre n'est encore qu'une critique préliminaire de l'intelligence artificielle ; je démonte la machine pièce par pièce, à la façon d'un ingénieur soucieux de démontrer que les « prouesses » techniques de nos algorithmes comportent des zones d'ombres. Je n'ai toutefois pas les outils théoriques nécessaires pour comprendre ces zones d'ombre et les théoriser. C'est pourquoi, dans la deuxième partie de cette thèse, je m'efforcerai de reconstruire la machine de l'esprit humain pièce par pièce, en conservant les apports de Kant, mais en tenant compte également du progrès récent de la science. Seule cette théorie unifiée permettra de comprendre en profondeur les différences entre l'intelligence artificielle et l'intelligence humaine.

3.1 La Machine manque-t-elle d'intuition spatiale ?

On entend souvent dire qu'une machine ne pense pas par elle-même ; qu'elle ne fait que répéter les réponses qu'on lui a savamment enseignées. Elle ne ferait pas preuve d'imagination, de flexibilité, n'aurait pas la capacité de comprendre ses propres processus cognitifs de façon métaréflexive. Dans ce scénario pessimiste, c'est la capacité même du jugement de la machine qui est remise en cause. L'on peut en effet soutenir que les algorithmes sont incapables de construire le concept d'objet en ramenant la diversité de la sensation sensible à une unité catégorielle. Dans le langage de Kant, on dira qu'il lui manque une fonction hénologique ; ou encore, que la machine ne peut pas subsumer les données sous un concept. Cette incapacité peut, à son tour, tenir au moins d'une de ces trois choses : un manque de sensation (pas de jugement synthétique), un manque de raison (pas de jugement analytique) ou l'absence de formes *a priori*.

L'option la plus radicale, et pas forcément la moins plausible, consiste à

dire de la machine qu'il lui manque la sensation proprement dite. La façon qu'elle a de recevoir les données et de les traiter, bien que s'apparentant au fonctionnement de l'esprit humain (ou toutefois à l'idée que nous nous en faisons), n'est rien qu'une imitation efficace. Cela signifierait que le fonctionnement de nos organes sensoriels, en dernier ressort, n'aurait rien de commun avec le traitement, quelque peu simplifié et réducteur, de l'ordinateur muni de capteurs. C'est une approche digne d'être considérée, dans la mesure où elle remet en question un postulat relativement peu contesté dans les sciences cognitives (ou du moins pas suffisamment), d'après lequel la perception est essentiellement un traitement électrochimique de l'information comparable à la manipulation, à chaque unité de temps, de séquences de « 0 » et de « 1 ». Gerald Edelman, prix Nobel de physiologie, n'hésite pas à rappeler à ses contemporains toute la naïveté philosophique cachée derrière une telle assimilation⁵ :

Dans de nombreux cercles scientifiques, la croyance selon laquelle le cerveau est un ordinateur reste largement répandue. Cette croyance est erronée pour plusieurs raisons. Premièrement, l'ordinateur fonctionne en utilisant la logique et l'arithmétique dans des intervalles très courts régulés par une horloge. Comme nous le verrons, le cerveau ne fonctionne pas selon des règles logiques. Pour fonctionner, un ordinateur doit recevoir des signaux d'entrée non ambigus.

Il y a dans ce passage un mot qui mérite, je crois, toute notre attention : c'est celui d'ambiguïté. L'ordinateur reçoit des informations exactes. Même lorsqu'une image floue lui est communiquée, il transforme cette image en matrice de nombres. Ces nombres ne présentent pas de propriétés spatiales immédiates. Le point situé dans le coin inférieur gauche n'est pas *a priori* plus éloigné du point inférieur droit. Il est certes possible de doter la machine d'une métrique, et de lui demander de calculer une distance, mais cette métrique est un choix du programmeur. Les seules différences réelles intrinsèques entre les points sont données par des valeurs numériques : trois nombres pour la couleur (rouge, vert, bleu), deux correspondant à la position du point sur la photo (abscisses et ordonnées), un nombre pour la luminosité, un pour l'albédo, etc.

Chez l'être humain, au contraire, il se pourrait que la sensation s'exerce au moyen des formes *a priori* de l'espace et du temps. Ces formes, dont l'existence reste bien sûr hypothétique, se manifestent en ceci que, même si la photo que nous regardons est plate, nous observons des perspectives. Nous distinguons l'avant-plan de l'arrière-plan, nous voyons des profondeurs apparaître, nous imaginons des trajectoires dans l'espace en trois dimensions,

5. GERALD EDELMAN, *Second Nature : Brain Science and Human Knowledge*, chapitre II : *Consciousness, Body and Brain*. La traduction est de moi.

etc. Si par exemple une forme est présentée dans l'arrière-plan, nous l'agrandissons inconsciemment dans notre perception pour tenir compte de l'effet de perspective. Cela veut dire que notre information n'est pas exacte, mais qu'elle est influencée par le contexte.

Ce processus inconscient est rendu tout particulièrement patent dans l'illusion de Ponzo (voir figure 3.4). Dans cette dernière, il vous est demandé de comparer deux barres horizontales et de déterminer — sempiternelle question — laquelle est la plus longue. Elles ont la même taille, mais les lignes de fuite dans l'arrière-plan créent l'illusion d'une perspective, si bien que la barre supérieure semble plus longue. Tout se passe donc comme si l'esprit humain projetait, malgré lui, ce qu'il voyait dans une perspective en trois dimensions. Le point intéressant, c'est que l'illusion apparaît même lorsque l'on sait que les barres ont la même taille.

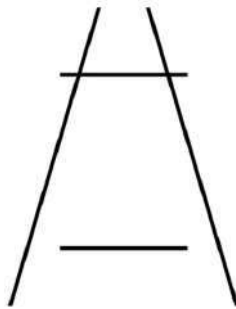


FIGURE 3.4 – Illusion de Ponzo : Les deux lignes horizontales ont la même taille, mais celle en haut de l'image semble plus grande.

Au sein des algorithmes, il n'est pas certain qu'une forme *a priori* de l'espace existe. Face à l'illusion de Ponzo, le programme se contentera d'utiliser sa métrique pour mesurer la longueur exacte des deux segments de droites. Il ne se trompera donc jamais. Il peut aussi mesurer quelle serait la différence de taille si ces lignes appartenaient effectivement à un espace à trois dimensions. Il lui suffirait pour cela d'utiliser les lois de la perspective. Pour lui, il serait parfaitement indifférent d'utiliser la première approche ou la seconde. Il n'a pas de regard privilégié sur les images qu'il analyse.

Il est certes possible d'entraîner l'ordinateur à répondre comme un humain à la question : « Quelle barre est la plus grande d'après toi ? ». Il suffit pour ce faire de lui donner un grand échantillon d'images, accompagnées des réponses de nombreuses personnes humaines, et de lui demander de construire un modèle statistique associant les bonnes réponses aux images visionnées. Il ne semble toutefois pas que cette illusion à laquelle il prétendra être soumis soit vécue comme une nécessité intrinsèque, qui témoigne d'une façon particulière de structurer sa perception du monde. Il se comporte à la façon du perroquet à qui l'on a appris à dire qu'il aimait lire Shakespeare.

Ce qui est extraordinaire avec l'illusion de Ponzo, et qui semble placer la perception humaine dans un autre registre que l'analyse statistique des points de l'image, tient à deux facteurs : (1) la prévalence du phénomène et (2) son inéluctabilité. Par *prévalence*, j'entends que l'effet de perspective apparaît même face à des images abstraites, obtenues par la disposition quasi aléatoire, sur une page vierge, de segments rectilignes. Il faut certes que ces droites soient placées de telle sorte à respecter les lois de la perspective, mais notre regard est si habile à dénicher des perspectives qu'il ne faudra pas beaucoup d'effort au dessinateur. Tout se passe comme si l'être humain cherchait absolument à faire entrer tout ce qui lui tombe sous les yeux dans une intuition spatiale. Il ne peut pas ne pas le faire.

Ceci nous reconduit devant le point (2), qui est *l'inéluctabilité* du phénomène. Même après avoir réalisé son erreur dans les grandes largeurs, il demeure rigoureusement impossible pour l'être humain de ne pas être victime de l'effet d'optique. Il peut, dans un dernier effort pour se débarrasser du problème, retourner l'image à 180 degrés, dans quel cas l'illusion s'amenuise en effet fortement. Mais sitôt qu'elle est replacée dans sa position initiale, elle réapparaît aussi sûrement qu'une pierre, lâchée du sommet d'un précipice, ne tombera vers la Terre. Si l'illusion était liée à une simple méprise au niveau du jugement rationnel, similaire à une erreur de calcul, il suffirait d'en prendre conscience pour la corriger. Or ici, tout se passe comme si la racine du mal était antérieure à toute perception possible.

Nous en venons ainsi à une importante remarque méthodologique : afin de déterminer l'existence de formes *a priori* dans la pensée d'un être humain ou d'une machine, il peut être plus instructif de s'intéresser, non pas à ce qu'ils sont capables de faire, mais à ce dont ils se montrent fondamentalement incapables. J'appelle ce principe le *principe soustractif d'intelligibilité*.

Définition 1 (Principe soustractif d'intelligibilité). *C'est en étudiant ce dont l'esprit se montre incapable par aucun effort conscient, et non ce dont il est capable, que l'on peut identifier la présence d'une structure a priori.*

En me fondant sur ce principe, j'aimerais maintenant me demander s'il existe, de façon analogue à ce que j'ai dit pour l'espace, un déficit dans la perception humaine qui témoignerait de l'existence d'une forme *a priori* du temps? Dans la théorie kantienne, en effet, l'intuition pure de l'espace s'accompagne d'une intuition pure du temps, qui concerne non pas le sens externe, mais les phénomènes internes aux sujets. C'est d'ailleurs le temps, on l'a vu dans la section [2.4.4](#) relative au schématisme, qui est à l'origine du jugement permettant de réunir diverses sensations sous une même unité conceptuelle. Il joue donc, pour ainsi dire, un rôle plus fondamental que l'espace. Son analyse, du reste, s'avère plus complexe.

3.2 La Machine manque-t-elle d'intuition temporelle ?

J'ai montré, dans le chapitre précédent, que la machine ne dispose pas de la même structure *a priori* de l'espace, mais j'ai laissé dans l'ombre la question de savoir ce qu'il en était pour la conception du temps. En guise de première observation, notons que si la perception humaine du temps était parfaitement objective, les événements se succéderaient avec la même rigueur qu'un métronome. Nous pourrions compter les secondes qui s'écoulent entre deux flashes lumineux, deux bruits, deux mouvements, sans être sujets à l'erreur ; les phénomènes de la conscience se succéderaient de façon linéaire et homogène. Peut-être que c'est le cas pour certains individus de fiction, comme le Funès de la nouvelle de Borges, qui est capable de donner le temps exact à n'importe quel moment. Qu'il coure, qu'il parle, qu'il saute, il se comporte comme une horloge ambulante⁶. Kant lui-même était assez proche de cette caricature, si l'on en croit l'écrivain Thomas de Quincey⁷, qui rapporte que les voisins du district réglaient leur montre sur ses ballades journalières.

Or dans la réalité, il se trouve que certaines illusions perceptives remettent en jeu cette conception. La plus connue est sans doute la chronostase, aussi désignée sous le nom de l'illusion d'horloge arrêtée. Cette illusion se manifeste par la sensation que le temps s'étend, qu'il dure plus longtemps qu'il ne devrait. Elle est vécue lorsque, après avoir écouté le "tic-tac" régulier de l'horloge en contemplant le paysage par la fenêtre, on détourne soudain les yeux pour les poser sur le cadran de l'horloge. On peut alors éprouver la sensation que l'aiguille s'immobilise imperceptiblement, qu'elle reste un peu trop longtemps dans la position qu'elle occupait au moment de la surprise du regard. Après ce bref temps mort, qui n'est pas sans évoquer un bug dans la matrice complexe de l'univers, l'aiguille semble reprendre sa marche implacable.

J'aimerais discuter cette illusion perceptive afin de déterminer si, oui ou non, l'esprit humain diffère du comportement d'une machine digitale. Pour mener à bien cette entreprise, je vais commencer par survoler les explications neuroscientifiques généralement admises à propos de la chronostase. Vers le milieu du *XX^{ème}* siècle, lorsque les chercheurs se sont intéressés pour la première fois à ce phénomène, ils avaient en effet la conviction, souvent implicite, que le cerveau se comportait à la façon d'une machine de traitement de l'information. Cette machine reçoit des images du monde extérieur, les modifie à satiété, avant de les envoyer dans le « théâtre de la conscience ». Or au début du *XXI^{ème}* siècle, des expériences viennent remettre en doute

6. BORGES, *Funès ou de la mémoire*

7. Thomas de Quincey, *Les derniers jours d'Emmanuel Kant*, éditions Mille et une nuits, 1996, p. 27-33.

ce paradigme computationnaliste, si bien que le mécanisme de la chronostase reste, aujourd'hui, globalement non élucidé. On a certes des indices, mais pas de consensus.

Dans ce chapitre, je vais suggérer que si l'on adopte un point de vue kantien sur le phénomène, en abandonnant l'idée que le cerveau se comporte comme un ordinateur, on peut obtenir un éclairage intéressant sur ce problème. Cette analyse que je vais produire visera à défendre l'idée que le temps n'est pas perçu de façon objective, comme lorsqu'on cherche à le mesurer avec une horloge, mais qu'il se présente dans notre conscience comme un tout où le présent, le passé et le futur immédiats se confondent. Ce n'est qu'à l'issue d'un découpage que ce vécu temporel amalgamé est mis à plat sur une ligne du temps possédant une orientation. J'insisterai sur l'idée que ce découpage est un *a priori*, une structure de l'esprit humain. Il opère à la manière d'un filtre. Les phénomènes, lorsqu'ils passent par ce filtre, prennent l'apparence d'une objectivité temporelle; mais il existe des cas précis où cette construction s'évapore comme une illusion.

3.2.1 Le Modèle des saccades

La chronostase a toujours fasciné l'observation scientifique, suscitant dans le chef des chercheurs les plus distingués des hypothèses variées. Jusqu'à la fin des années 90, l'explication dominante consistait à imputer cette dilatation du temps subjectif aux mouvements rapides des yeux, appelés saccades. Il est en effet connu que, lorsque les yeux bougent brièvement, votre vision se brouille pendant une fraction de seconde. Pour compenser cet effet, le cerveau corrige la perception en substituant, aux images floues que vous voyez, les images non floues qui sont disponibles avant ou *après* la saccade. Vous avez bien lu : le cerveau remplace l'image floue vue à un instant t par une image nette vue l'instant *d'après*. Il s'agit donc d'une correction visuo-spatiale *a posteriori*.

Je vais préciser le fonctionnement exact de ce mécanisme neuro-oculaire. Avant de me lancer dans les détails techniques, j'aimerais préciser que ce phénomène de saccades dont je parle est quelque chose dont vous faites chaque jour l'expérience, même lorsque vous vous tenez au repos. Vous n'avez sans doute pas conscience des micromouvements de votre œil qui s'opèrent en cet instant même, tandis que vous lisez ces pages, et pourtant — comme dirait l'autre — et pourtant il tourne. Votre œil, loin de se tenir sagement tranquille, oscille sans cesse dans vos orbites. Tout se passe ainsi comme si votre perception apparemment fluide du monde, ce long fleuve tranquille et cinématographique qui vous tient lieu de vision, était en réalité un torrent cahoteux, capricieux et incertain, jonché d'innombrables micro-corrections.

Le cadre épistémologique qui préside à ce mécanisme des saccades est analogue à une salle de cinéma. Votre cerveau est une machine savante, qui reçoit des images, qui fait des calculs et opère des corrections sur ces images

de façon à rendre, à son patron la conscience, une vision du monde la plus propre, la plus esthétique, la plus pratique que possible. Lorsqu'il fait bien son travail, les images sont positionnées sur une longue bande de lecture, et lues les unes après les autres. La conscience, confortablement assise sur son siège, observe le résultat final, sans avoir conscience des constructions qui ont eu lieu en coulisses. Le film de la conscience se déroule sans fin, toujours à la même vitesse. Une image gravée au début de la bande sera perçue comme ayant eu lieu avant une image gravée à la fin de cette bande.

Ce modèle que je vous décris — notons-le d'emblée — n'est pas exactement celui manipulé par les neuroscientifiques. Je pense néanmoins qu'il capture correctement les intuitions épistémologiques sous-jacentes aux explications neuroscientifiques de la chronostase. Lors d'une saccade, tout se passe comme si le cerveau recevait un échantillon d'images endommagées. La question qui se pose est de savoir ce qu'il doit en faire ? S'il les grave sur la bande de lecture dans l'ordre où il les reçoit, la conscience risque d'être déboussolée. C'est pourquoi il préfère jeter les images floues et faire des copies de celles qui sont encore nettes.

Pour résumer la situation, il y a donc quatre événements qui surviennent sur le ruban de la conscience lors d'une saccade : (1) celui où vous commencez la saccade, (2) celui où vos yeux sont en déplacement, (3) celui où vous vous arrêtez sur l'horloge et (4) celui où vous voyez l'aiguille bouger. Il se trouve que, entre les événements (2) et (3), votre vision est floue à cause du mouvement oculaire, si bien que le ruban présente des taches dignes de l'art abstrait⁸. D'après le modèle des saccades, il y a donc une compensation à effectuer par le cerveau. Ce dernier va remplacer, *a posteriori*, la partie du ruban qui est floue par une image nette. Or, la seule image dont il dispose, c'est celle que vous voyez à l'instant (3). Vous avez donc l'impression que l'image de l'instant (3) commence bien avant l'instant (3) et peu après l'instant (2). Cela vous donne la sensation, *a posteriori*, que l'horloge s'est arrêtée plus longtemps qu'elle ne devrait.

3.2.2 Objection : la Chronostase auditive

L'explication de la chronostase que j'ai exposée ci-dessus repose sur un postulat : celui d'après lequel la chronostase est toujours concomitante à une saccade. Or l'expérience quotidienne semble contredire ce préjugé. La chronostase peut survenir dans des contextes non visuels, comme lors d'une conversation téléphonique. Si vous faites passer le combiné d'une oreille vers l'autre, afin de vous emparer de votre agenda, vous aurez l'impression, au moment de reprendre le fil de la conversation, que l'intervalle de temps

8. Le lecteur ne constatera sans doute pas cet effet de flou lorsqu'il bouge la tête. D'un point de vue neurologique, le flou est expliqué par une baisse de la sensibilité visuelle survenant pendant la saccade. Cet effet est notamment étudié par Knoll, Jonas, *Spatio-temporal representations during eye movements and their neuronal correlates*, 2012.

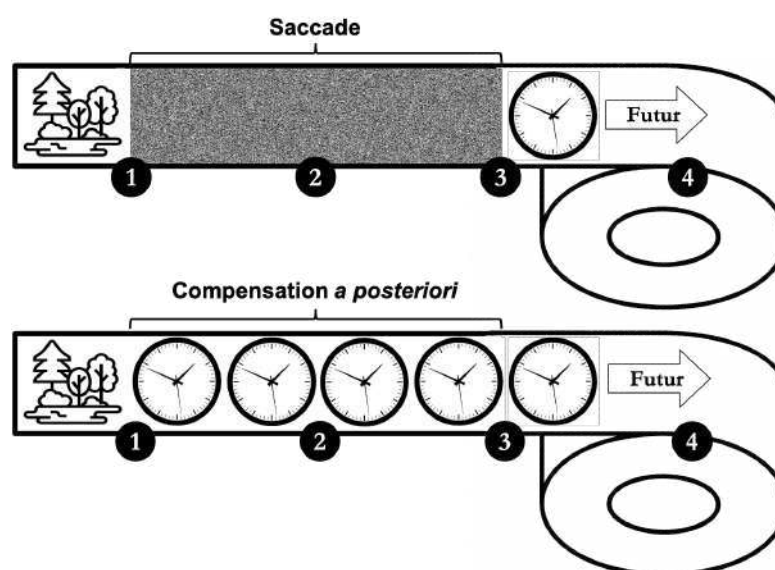


FIGURE 3.5 – Dans le modèle physiologique de la chronostase, le temps est une structure objective comparable à un ruban qui s’imprègne de nos perceptions sensorielles. Lorsqu’un mouvement de saccade a lieu, le cerveau remplace l’image floue qui en résulte par la première image nette observée à l’issue de cette saccade. Cet effet compensatoire est *a posteriori*.

correspondant à cette brève interruption est bien plus long que vous ne le pensiez. Cela vous donne l’étrange conviction que votre interlocuteur, devinant par quelque mystérieux procédé que vous ne pouviez pas l’écouter aussi attentivement à cet instant-là, s’est interrompu pour vous attendre. Vous avez la nette impression qu’il reste silencieux ou que la ligne a été coupée.

Cette forme auditive de la chronostase a reçu des preuves expérimentales au tout début des années 2000. Dans une étude menée à l’université d’Oxford, un groupe de sujets devait mettre un casque audio sur la tête et évaluer le temps s’écoulant entre deux stimuli sonores⁹. Cette évaluation avait lieu en leur demandant de cliquer sur une touche dès qu’un beep retentissait et dès qu’il s’arrêtait. L’objectif consistait à mesurer un décalage entre le temps objectif du beep enregistré par l’ordinateur et le temps subjectif mesuré par l’individu, de sorte à identifier précisément les conditions dans lesquelles ce décalage apparaît.

Dans la condition de contrôle, quatre beeps consécutifs étaient émis dans l’oreille droite, séparés chacun d’une seconde. Cette première séquence était suivie d’un temps d’attente variable, après quoi quatre autres beeps reten-

9. IONA HODINOTT-HILL, KAI V THILO, ALAN COWEY, VINCENT WALSH, *Auditory Chronostasis : Hanging on the Telephone*, 15 octobre 2002, pages 1779-1781.

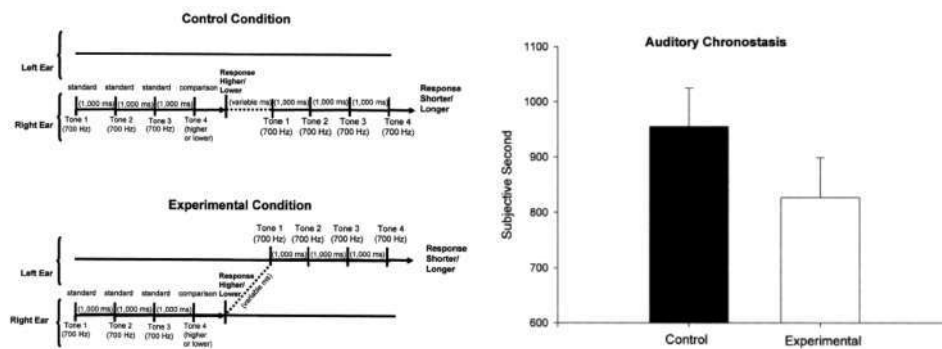


FIGURE 3.6 – Dans cette expérience réalisée à Oxford, les sujets doivent déterminer le temps qui s’écoule entre deux beeps sonores. Dans la condition contrôle, tous les stimuli sont émis dans l’oreille droite (figure en haut à gauche). Dans la condition expérimentale, les quatre premiers beep sont émis dans l’oreille droite et les quatre suivants dans l’oreille gauche (figure en bas à gauche). Ce changement d’attention induit une modification de la perception du temps (figure de droite).

tissaient, à cette même fréquence d’un son par seconde. Les participants devaient dire si l’intervalle de temps séparant les quatre premiers beeps était de même longueur que l’intervalle séparant les quatre suivants. Pour s’assurer qu’il n’y ait pas de méprise, le dernier des quatre premiers beeps était volontairement émis à une fréquence plus basse, de sorte à signaler aux participants que les beeps suivants appartenaient à la deuxième série de sons. Dans cette condition de contrôle, les sujets n’étaient pas victimes de distorsion de perception temporelle. Leurs évaluations des intervalles séparant chacun des quatre beeps au sein des deux séquences sonores convergaient autour d’une seconde.

Dans la condition de test, on faisait varier le dispositif de sorte à produire l’illusion de la chronostase. Les quatre premiers beeps étaient émis dans l’oreille *droite*, à un intervalle d’un son par seconde, et les quatre suivants dans l’oreille *gauche* (toujours à la même fréquence). Ce nouveau protocole visait à modéliser, en laboratoire, les situations de vie où un individu faisait passer son téléphone d’une oreille vers l’autre. Les sujets étaient préalablement informés que le son allait être transmis dans l’autre oreille, mais ils n’en étaient pas moins victimes de l’illusion temporelle : ils avaient l’impression que les intervalles de temps séparant les stimuli sonores au sein de la seconde séquence se rallongeaient.

Ce résultat est conforme à l’effet attendu en cas de chronostase. L’expérience d’Oxford suggère ainsi que l’illusion d’optique, contrairement à ce que l’on pensait jusque-là, ne s’explique pas, ou pas seulement, par le mécanisme

physiologique des saccades, mais appartient à un phénomène plus large. Plus fondamentalement, il semblerait qu'il faille réviser le modèle d'après lequel il existe une horloge interne au sein de l'organisme. Le temps ne se présente pas comme un ruban qui attend de recevoir la marque de nos différentes perceptions. Comment, dès lors, peut-on expliquer la perception du temps ? Existe-t-il une interprétation de la chronostase compatible avec l'idée kantienne qu'il existe, dans l'esprit humain, une intuition *a priori* du temps ?

3.2.3 Hypothèse : perception différentielle

Depuis que l'expérience d'Oxford a invalidé l'explication par le mécanisme des saccades, les scientifiques ont tenté de trouver des modèles alternatifs. Plusieurs pistes de réflexion sont suggérées dans la littérature, qui s'orientent autour de trois grands axes méthodologiques :

- **La Motricité** : Certains auteurs continuent d'affirmer que, même si la chronostase se manifeste en dehors des saccades oculaires, le phénomène reste lié à un mouvement de l'organisme. Dans l'expérience d'Oxford que j'ai rapportée, les sujets doivent en effet bouger leur index pour cliquer sur un bouton.
- **L'Attention** : Certains chercheurs affirment que la cause de la chronostase pourrait aussi être liée aux changements attentionnels. Ce n'est pas le mouvement d'un organe qui serait alors déterminant, mais le fait que notre concentration passe d'un endroit à l'autre de la pièce ou d'une oreille à l'autre^[10].
- **Le Changement** : Depuis les années cinquante, des études montrent que la perception du temps est influencée par la nature du stimulus lui-même^[11]. On identifie en particulier que les contrastes au sein des stimuli semblent accompagner l'apparition de l'illusion temporelle.

Quoique chacune de ces hypothèses scientifiques constitue une voie prometteuse pour l'élucidation des mécanismes physiologiques exacts de la chronostase, elles ne révèlent pas forcément une structure *a priori* de la perception du temps. Ces hypothèses ne font en effet que dresser des corrélations

10. Dans une expérience, Rose, D., Summers, J. produisent des séquences de flash lumineux et observent la présence de chronostase lors de l'apparition du premier flash et, parfois, lors de sa disparition. Les sujets ne bougent pas la tête ni les yeux, ce qui suggère que ce n'est pas la saccade qui est responsable du phénomène. Voir *Duration illusions in a train of visual stimuli*, Scopus, Volume 24, Issue 10, 1995, Pages 1177-1187.

11. Voir notamment ROELOFS C.O.Z., ZEEMAN W.P.C., *Influence of different sequences of optical stimuli on the estimation of duration of a given interval of time*, Acta Psychologica. 1951. Consulter aussi SCHIFFMAN H.R., BOBKO D. J., *Effects of stimulus complexity on the perception of brief temporal intervals*, J Exp Psychol, 1974.

a posteriori entre différentes variables empiriques. Toute la méthodologie expérimentale consiste à observer comment la modification d'un facteur tel que l'attention va induire une hausse ou une baisse de l'illusion perceptive — et dans certains cas supprimer la présence de cette dernière. Or d'après le principe d'intelligibilité soustractive que j'ai défendu dans le chapitre précédent, l'existence d'une forme *a priori* de la perception ne peut être révélée que par la présence d'une illusion qui résiste à tout *effort conscient*.

La méthode à adopter, dans le cadre de ma démarche, consiste donc à déterminer, parmi les quatre hypothèses scientifiques qui ont été soulevées, quelles sont celles qui ne peuvent pas être contrôlées par un *effort conscient* ? Il semble possible de contrôler son attention sur un effort suffisant de la volonté, de bouger ou non la tête, les yeux, les mains ; il est en revanche rigoureusement impossible de contrôler le changement qui existe entre deux perceptions. Quelle que soit la façon dont on tourne le problème, lorsque votre tête va de la contemplation du paysage à l'observation de l'aiguille sur le cadran, ou lorsque votre téléphone passe de l'oreille gauche à l'oreille droite, il y a *de facto* un changement dans le contenu de votre perception — sur lequel vous n'avez aucune emprise.

Dans la mesure où mon objectif est de déterminer s'il y a une structure *a priori* du temps, et sous quelle condition elle se manifeste, j'aimerais pousser plus avant cette hypothèse selon laquelle la chronostase est liée au changement, de sorte à lui donner une interprétation philosophique. On a vu, dans le cas de l'espace, que l'esprit humain ne semble pas appréhender les points de l'espace de façon isolée ; ces derniers n'ont de réalité, pour lui, que dans les rapports qui les unissent aux autres points de l'espace. De la même manière, j'aimerais suggérer que les instants n'existent pas dans la perception humaine de façon individuelle. Tout ce que notre esprit peut observer, c'est le changement existant entre deux événements.

Pour donner à cette idée que je formule une analogie éclairante, on peut penser aux notions de vitesse et de position dans le cadre des trajectoires physiques. Lorsqu'un corps se déplace dans l'espace, la conception naïve consiste à dire qu'il occupe, à chaque instant, une position précise avec une vitesse précise. Si toutefois les instants n'existaient pas, qu'est-ce que cela donnerait ? Même pour un objet macroscopique tel qu'un ballon, il n'est pas évident de mesurer avec exactitude la position ou la vitesse à un instant donné. Si vous prenez une photo de la trajectoire, le ballon va vous apparaître flou, si bien qu'il vous faut changer l'obturation de votre appareil. Avec les réglages adéquats, vous pouvez espérer avoir la position exacte du ballon, mais sur le cliché ainsi obtenu, il n'y a plus trace de la vitesse.

La vitesse d'un corps en déplacement n'est en effet pas une propriété intrinsèque de celui-ci, mais elle est définie de façon différentielle, comme étant le rapport entre la distance parcourue et le temps nécessaire pour effectuer cette distance. Si la balle prend une seconde pour aller de A vers B, et que ces points sont éloignés de vingt mètres, elle fera du vingt mètres

par seconde. Il ne s'agit donc pas d'une vitesse instantanée, mais de la vitesse moyenne sur l'ensemble de cette trajectoire.

Pour connaître la vitesse à un moment plus précis de ce mouvement, vous devez rapprocher les points A et B, mesurer combien de temps le ballon a pris pour parcourir, non pas vingt mètres, mais vingt centimètres, puis vingt millimètres, vingt nanomètres. En rapprochant toujours plus la distance entre A et B, on tend vers une approximation tendancielle exacte de la vitesse, que l'on nomme la vitesse instantanée. Cette vitesse est cependant une abstraction mathématique. Il n'est pas du tout évident qu'une telle grandeur existe autrement que dans la fiction de nos théories. Le principe de Heisenberg semble d'ailleurs indiquer que, quand vous arrivez à des échelles de grandeurs quantiques, cet idéal de la vitesse ou de la position instantanée s'effondre.

Ce que je souhaite illustrer avec cet exemple, c'est que les contenus perceptifs de la conscience ne sont pas forcément des grandeurs absolues, qui existent à chaque instant, mais qu'ils dépendent des autres événements ayant lieu dans le voisinage temporel immédiat. Cette notion de voisinage est gardée volontairement floue pour le moment. Ce que je veux exprimer, en substance, c'est que l'être humain n'appréhende pas les événements de façon séquentielle, en faisant entrer les instants en file indienne par la porte de sa conscience. Ce grand fleuve qu'est le temps n'est pas composé de gouttes d'eau indivisibles, identiques à des billes de billard. Ce n'est pas le ruban d'un grand film photographique destiné à égayer le « théâtre de la conscience ». Le temps subjectif se présente au contraire comme un tout que notre esprit transforme en une séquence chronologique ayant un début et une fin.

Cette homogénéité, cette totalité sous laquelle le temps se présente est peut-être à l'origine du vertige, de l'hébétude, de l'aveu d'impuissance qui nous saisit lorsque nous tentons d'appréhender l'instant présent ? Un tel instant ne semble pas exister, ou du moins résiste-t-il aux spéculations philosophiques les plus attentives. Avant que nous le pensions, c'est un futur ; et dès lors que nous le pensons, c'est un passé — si bien que l'instant, le passé et le futur immédiats sont intriqués l'un dans l'autre. Ils se répondent les uns les autres dans un éternel jeu de miroirs.

Au lieu de concevoir le temps comme un ruban infini, que l'esprit balaye de façon linéaire, j'aimerais dès lors proposer une autre représentation géométrique du temps. Celle-ci ne représente, rigoureusement parlant, qu'une analogie ; je pense néanmoins qu'elle peut s'avérer éclairante. Dans la mesure où nous ne percevons pas les instants en soi, mais les changements se produisant dans un voisinage temporel, ce voisinage pourrait être assimilé à une boule de faible rayon. Notre perception de ces phénomènes amalgamés consisterait alors à projeter cette boule sur une droite temporelle allant du passé vers le futur. Selon l'orientation que vous donnez à la boule avant la projection, il est possible que deux événements A et B occupent la même position sur la ligne du temps. La simultanéité n'est donc pas un absolu

(voir figure 3.7).

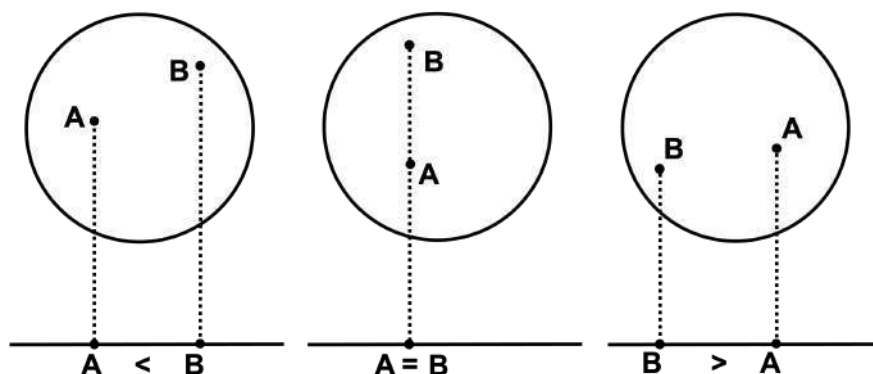


FIGURE 3.7 – Lorsque l’on projette deux points A et B d’une boule sur une droite, il est possible de l’orienter de telle sorte que (i) A soit avant B, que (ii) A et B soient confondus ou que (iii) B soit avant A.

Ce phénomène dont je parle est connu des ingénieurs du son. Un important problème qui se présente à eux lors d’une projection ou d’un montage consiste en effet à synchroniser l’audio avec l’image. À leur plus grand bonheur, il se trouve qu’un décalage de cent millisecondes ne sera pas aperçu par un spectateur. La cause en est que l’information visuelle et l’information auditive ne sont pas traitées à la même vitesse par le cerveau. Même lorsque nous observons des personnes marcher dans la rue, il y a toujours un décalage. Le cerveau construit donc la temporalité des événements *a posteriori*, en se basant sur différentes informations causales. Un son de chute doit être causé par l’objet qui tombe sur le sol. Une correction est effectuée.

Une attestation scientifique de cette idée peut être trouvée dans une étude menée par l’University College London, où les participants doivent appuyer sur un bouton pour allumer la lumière¹². Le dispositif est construit de telle sorte qu’un délai de 100 millisecondes ait lieu entre le moment où ils cliquent et le moment où la lumière apparaît. Au départ, ils ressentent ce décalage, mais après un certain temps d’adaptation, ils finissent par voir les deux événements simultanément. Si, à cet instant, on synchronise réellement le bouton et la lumière, les sujets ont alors l’impression que la lumière apparaît *avant* qu’ils n’appuient sur le bouton — autrement dit, il y a une interversion de l’action et de la sensation. Les auteurs de l’étude font remarquer que des effets de désynchronisation similaires existent également lorsque le participant est passif et ne doit pas agir en appuyant sur un bouton, mais que l’intensité du phénomène est alors 50% moins grande.

12. HAGGARD P, CLARK S, KALOGERAS J., *Voluntary action and conscious awareness*, Nature Neurosciences 2002.

3.2.4 Temps subjectif vs temps objectif

Le phénomène de la chronostase, dont j'ai fait l'analyse, supporte donc, ou est du moins compatible, avec l'interprétation qu'il existe une structure perceptive du temps, chez l'être humain, qui n'est pas comparable au traitement d'une machine automatique. Les études dont j'ai parlé semblent confirmer, en particulier, l'idée que la simultanéité et la succession des événements ne sont pas des réalités absolues, mais des constructions. Il me reste toutefois à insister sur un point important : cette construction perceptive que j'ai en vue n'est pas réalisée *a posteriori*, mais agit bien plutôt comme un filtre *a priori*.

Ce que j'entends par là, c'est que le cerveau ne prend pas simplement les événements incidents, comme un son ou une image, pour les piquer un à un sur sa ligne du temps. Dire cela, c'est penser que le temps est une réalité objective du monde, un axe tracé à l'encre indélébile dans l'Univers, que notre cerveau tente de reconstruire. Cette vision linéaire du temps, héritée de Newton, est naïve. Elle ne tient pas compte des progrès récents de la physique¹³. Par ailleurs, le grand enseignement de Kant, je l'ai exprimé à maintes reprises, consiste à insister sur le fait que ce sont les objets qui se calquent sur nos structures du temps et de l'espace, et non l'inverse. Il ne faut donc pas supposer que le temps existe comme une réalité fondamentale du monde, mais il faut admettre l'idée que le temps objectif et subjectif sont, l'un comme l'autre, des structures de notre pensée.

La chronostase n'est donc pas — comme on le dit souvent — une déformation de la réalité objective du temps qui serait imputable à notre subjectivité. Elle est plutôt l'indice que ni le temps objectif ni le temps subjectif ne représentent adéquatement les événements du monde. Qu'est-ce que je veux dire par là exactement ? Lorsque vous dessinez des objets en trois dimensions, comme le globe terrestre, sur une feuille plate, il y a des déformations qui se produisent. C'est ainsi que la projection de Mercator change le volume des continents, tout en conservant les angles, tandis que la projection de Peters sacrifie les angles pour conserver les superficies. De la même manière, le temps objectif et le temps subjectif sont deux façons de projeter, dans un plan intelligible, les événements du monde.

Le temps objectif est obtenu en prenant un mouvement régulier comme référentiel. Il s'agira donc de se rapporter au mouvement d'une aiguille sur le cadran finement huilé de l'horloge ou aux radiations de l'atome de césium 133. Il n'est cependant pas certain que ces phénomènes soient parfaitement réguliers dans l'absolu. Ce ne sont finalement que des hypothèses, certes efficaces et commodes, mais invérifiables. Si le monde dans son entier ralentissait, que tous les mouvements s'effectuaient moins rapidement, serions-nous en mesure de le percevoir ?

13. Pour une revue critique et très abordable de la notion du temps en physique, se référer à l'ouvrage de PETER W. EVAN intitulé *A study of Time in Modern Physics*.

Le temps subjectif, quant à lui, est construit par l'esprit en prenant la mesure du changement ou du décalage existant entre deux perceptions. Lorsque vos yeux passent de la vision d'un paysage champêtre à l'observation de l'aiguille de votre montre, vous vivez plus de changements que lorsque vous voyez l'aiguille passer tranquillement d'une position à l'autre. Il y a en effet un plus grand contraste entre les champs de blé baignés de soleil et votre cadran métallique qu'il n'y en a lorsque vos yeux restent en scrutation sur les aiguilles. Le temps subjectif conserve pour ainsi dire — si l'on m'autorise à poursuivre cette analogie cartographique — le volume des événements ; là où le temps objectif conserve les distances.

Il est important de noter que c'est bien l'ensemble des contenus mentaux qu'il faut prendre en compte dans cette mesure du temps, et non seulement le champ visuel. Il se peut que vous regardiez l'aiguille, mais que votre pensée s'évade tant et si bien que le temps peut vous paraître interminable. Lorsque vous attendez le bus, le temps semble lent, car vous vous attendez en permanence à voir le bus surgir à l'angle de la rue. Vous regardez dans la direction de sa venue, vous l'imaginez, il est là sous vos yeux, prêt à vous accueillir ; puis vous réalisez que la rue demeure irréconciliablement vide. Il ne s'est pas écoulé une seconde que vous l'imaginez à nouveau. Ce contraste rapide entre l'idée que vous vous faites (votre désir) et l'image de la réalité crée un tel contraste que le temps vécu est bien supérieur au seul mouvement de l'aiguille.

Le fait que le temps objectif et subjectif soient les deux projections possibles d'une même réalité amalgamée ne signifie pas qu'ils soient d'égales valeurs. Au même titre que la projection de Mercator s'est imposée dans tous les planisphères — en raison de son efficacité maritime — la définition objective du temps s'impose par une plus grande précision de mesure. Il n'est cependant pas exclu que la représentation subjective du temps ait de l'intérêt dans certains contextes. Il pourrait par exemple y avoir un avantage évolutionnaire à ce que notre espèce ne soit pas dotée d'une vision objective du temps : le ralentissement temporel que nous éprouvons dans certaines situations émotionnelles, face à un danger imminent par exemple, pourrait être lié à notre capacité à traiter les informations, et leurs multiples connexions, comme un tout hyperconnecté (et non comme des sommes de stimuli individuels). L'objet de cette thèse est justement de comprendre les limitations rencontrées par les algorithmes, qui séquentialisent le temps et l'espace, lorsqu'il s'agit d'aborder des problèmes requérant une analyse plus abstraite, plus holistique, plus générale — comme les fameux exercices ARC.

Pour résumer ce qui a été dit, on pourrait supposer que l'esprit humain, quelle que soit l'information à laquelle il est confronté, ne peut voir cette information que dans une structure spatiale et temporelle préalable. Il ne voit pas les points de l'espace et du temps de façon isolée, comme tout autant de taches de couleur jetées à plat sur un ruban de papier. Il les conçoit au contraire de façon holistique. Les points acquièrent une intelligibilité dans

la disposition qu'ils occupent les uns par rapport aux autres ; c'est-à-dire dans la différence qui les sépare. Cette différence peut être entendue au sens spatial (une distance entre deux points), ou au sens temporel (un contraste entre deux événements).

À l'inverse de l'esprit humain, l'ordinateur prend les points isolés comme des données brutes, indépendamment des relations qui les lient les uns aux autres. Il n'est pas sujet aux illusions d'optique, mais il se montre incapable de relier efficacement ces points pour leur donner une signification. Son fonctionnement, purement séquentiel, le rend moins efficace à l'intelligence du tout. On pourrait dire qu'il répond d'une structure atomique du temps et de l'espace : les phénomènes sont séquencés et décomposés de façon à se présenter comme une conjonction d'informations disjointes les unes des autres.

Bien que ces observations constituent des critiques tout à fait raisonnables, il est peut-être extrême d'aller jusqu'à dire que la machine, sous prétexte qu'elle ne traite pas l'information d'après les mêmes structures spatio-temporelles que l'être humain, ne soit pas capable de jugement. Il est plausible de penser qu'il existe bel et bien un jugement dans la machine, entendu comme une capacité à manipuler des concepts, mais que ces concepts sont purement rationnels. En l'absence des formes pures de la sensation, qui tisse des relations intrinsèques entre les objets, l'algorithme doit se contenter de lier les idées les unes aux autres de manière logique. Dans le langage de Kant, cela revient à dire que la machine est un sujet pour lequel la déduction objective échoue. Par « déduction objective », rappelons-le, Kant cherche à comprendre comment il se fait que les catégories de l'Entendement puissent s'appliquer à quelque chose, plutôt que rien. En l'absence de sensation, ces catégories logiques existent, mais tournent à vide. C'est un moulin à moudre qui n'a pas de matière première à se mettre sous la dent !

3.3 La Machine manque-t-elle de raison contemplative ?

J'ai suggéré que la machine ne traitait pas les informations spatiales et temporelles d'après les mêmes structures *a priori* que l'être humain. Si l'on en croit cette ligne argumentative, cela signifierait qu'elle n'aurait pas de faculté sensitive à proprement parler. La sensation est en effet définie, chez Kant, comme le fait de se représenter les phénomènes dans un espace d'objectivité préalable. Lorsque nous voyons la couleur rouge, nous ne voyons pas les photons et les ondes telles qu'elles sont *en soi*. Les photons ont-ils seulement une forme connue ? Comment aurions-nous accès aux choses vraies, telles qu'elles existent présupposément au-delà de la représentation que nous nous en formons dans notre esprit ? Les corps physiques ne nous apparaissent jamais que sous des formes qui dépendent de nos structures perceptives préalables. C'est parce que l'intuition de l'être humain est spatiale et temporelle que tous les phénomènes physiques se présentent à nous dans le temps et dans l'espace. Il en est ainsi depuis que l'homme est homme. Et

il continuera d'en être ainsi, même après que l'on eût démontré que l'espace et le temps n'eussent pas de réalité physique en dehors de notre sensation subjective.

Il découle de cette analyse que la machine ne se représente pas les phénomènes sous une forme subjective, mais qu'elle les envisage uniquement d'après des termes de langage objectifs : des nombres, des symboles, des définitions déterminées par avance. On pourrait dire, en des termes imagés, que la réalité est mise à plat. Il manque une épaisseur ontologique, une différenciation entre les signes utilisés pour parler des objets et l'idée des objets eux-mêmes. Il s'ensuit que les algorithmes ne se rapportent pas à des phénomènes, mais à des symboles numériques. Ils traitent l'information de façon syntaxique, comme un ensemble de nombres objectifs et manipulables, sans détenir de sémantique.

Bien sûr, la question de savoir ce que sentirait un être artificiel, s'il venait jamais à déclarer éprouver des émotions, demanderait de se mettre *dans sa tête*, est un indécidable sur le plan métaphysique. Je n'ai pas en vue ici de dire quoi que ce soit à propos de la subjectivité d'une machine, mais je souligne seulement qu'il manque une certaine structure, dans la perception de l'information, qui puisse être assimilée, sur le plan fonctionnel, à ce que l'on connaît de la sensation humaine.

Après avoir discuté du problème de la sensation, la question qu'il me reste à examiner est celle de savoir si les calculs propres à la machine relèvent d'une rationalité au plein sens du terme, ou s'il s'agit d'autre chose encore ? Autrement dit, j'aimerais me demander dans quelle mesure l'on peut penser que les algorithmes sont dotés de raison. La posture qui consiste à dire que la machine n'a pas de raison, qu'elle ne fait tout au plus qu'imiter plus ou moins fidèlement la pensée logique, semble assez surprenante. Si l'on en croit le sens commun, il est généralement plus facilement admis que l'ordinateur manque de sensation que de réflexion abstraite. La raison de ceci vient de ce que les premières machines étaient des machines à calculer et que le calcul est souvent vu comme l'archétype de la rationalité.

Or, si l'on s'en tient à la théorie kantienne, la raison ne se définit pas *uniquement* par la manipulation de symboles, mais *principalement* par l'aptitude à remonter aux principes inconditionnels de la vérité. À ce titre, Kant est tributaire de la pensée de Platon, qui voyait dans la *noësis* cette aptitude à chercher l'anhypothétique : en partant de simples hypothèses, et en s'élevant de proche en proche, l'intellect cherche à capturer l'unité première du monde ; à comprendre ce qui se tient derrière, ou au-delà, de tous les principes. Cette unité, dans la *République* de Platon, c'est l'Idée du Bien, qui est souvent comparée au soleil. Cette idée est ce sans quoi aucune autre idée ne pourrait être conçue. Elle est telle la lumière qui permet aux yeux de voir le monde. C'est ce qui est dit, en substance, dans ce passage du livre VI de la *République* :

*Quel est donc de tous les dieux du ciel celui [...] dont la lumière permet à nos yeux de voir de la meilleure façon possible ? [...] C'est le soleil évidemment que tu me demandes de nommer. [...] Comme on a raison de penser que la lumière et la vue sont semblables au soleil, mais tort de croire qu'elles sont le soleil, de même, dans le monde intelligible, il est juste de penser que la science et la vérité sont l'une et l'autre semblables au bien, mais faux de croire que l'une ou l'autre soit le bien*¹⁴.

Dire de la machine qu'elle n'a pas de Raison ne revient évidemment pas à lui nier la capacité, purement factuelle, de produire des déductions à partir de symboles. C'est une évidence que les algorithmes, de par leurs structures mêmes, sont aptes à dériver des équations logiques. Ce que je voudrais contester ici ne concerne finalement qu'une dimension particulière de la raison, qui différencie celle-ci d'une pure et simple rationalité. La raison, en effet, n'est pas une faculté de calcul, mais c'est une certaine aptitude à la contemplation des idées. Il y a dans la vie intellectuelle une expérience de clarté, de lucidité, qui consiste à voir ses propres processus de pensée comme s'il s'agissait d'un objet du monde. Cette expérience est mêlée étroitement à un sentiment esthétique.

Les scientifiques qui ont traversé de tels états rapportent avoir senti quelque chose qui touche à une espèce de grâce intellectuelle. Entre les interstices d'une équation, c'est un pan de l'Univers qui, tout à coup, leur est révélé. Le mystère de l'Univers, l'immensité du tout, la structure, l'ordre, l'harmonie intrinsèque des formes géométriques, l'hébétude d'une idée qui semble dépasser, par son contenu, ce que l'esprit humain est capable de mettre en elle, est responsable d'un étourdissement métaphysique. L'essence du monde semble se dévoiler, l'espace de quelques instants, à celui qui parvient, au hasard d'une longue étude, à mettre le pied dans la clarté des concepts.

Cette expérience visuelle, quasi existentielle, est ce que l'on pourrait appeler, à proprement parler, l'intuition intellectuelle. Au même titre que l'intuition spatiale et temporelle donnent une forme aux phénomènes perceptifs, en prédéfinissant la manière dont un événement nous apparaît, l'intuition intellectuelle donnerait aux idées abstraites une forme essentielle, en nous permettant de saisir par l'intellect ce qui n'a pas de prise dans le monde sensible. Si cette position que j'évoque peut paraître mystique de prime abord, il suffit de penser à l'idée de l'infini. Il y a présence, dans l'esprit humain, de concepts dont l'essence dépasse l'entendement, mais dont on peut pourtant se faire une idée claire et distincte. Lorsque vous pensez à l'ensemble des nombres entiers naturels, vous seriez bien en peine de les visualiser dans leur ensemble. Si je vous dis que le nombre $2^{136.279.841} - 1$ est le plus grand nombre premier connu, vous serez sans doute pris d'un étour-

14. PLATON, *La République*, livre VI.

dissement en apprenant qu'il possède 41024320 chiffres. Cet étourdissement que l'on éprouve en tentant de se représenter des expressions aussi banales que « l'ensemble des nombres premiers » nous rapproche asymptotiquement d'une compréhension plus essentielle, plus rationnelle, plus intellectuelle, d'un infini irreprésentable au moyen de nos sens, qui confine à la contemplation esthétique.

Notons que cette idée d'intuition intellectuelle que j'examine ici ne s'appuie pas sur la théorie kantienne, mais sur le platonisme. Pour Kant en effet, l'intuition intellectuelle n'existe pas. Il ne saurait être plus clair à ce propos. Sa *Critique de la raison pure* est un plaidoyer en défaveur du fait que les idées aient une existence indépendante de la sensation. La raison pure, d'après lui, peut certes s'exprimer à propos des choses morales, mais elle ne peut pas conduire à une connaissance proprement scientifique. Kant proscriit la *noësis* de Platon, conçue comme la recherche de l'anhypothétique.

Du point de vue de Kant, donc, il n'existe de toute façon pas d'intuition intellectuelle spéculative dans le jugement humain. Les concepts de Dieu, de l'Univers, de l'infini se ramènent à des contradictions. En schématisant quelque peu, on peut dire de ces concepts que ce sont des symboles similaires à ceux manipulés par une machine sans âme. On peut certes établir des règles sur la façon d'utiliser ces symboles, tout comme on peut utiliser des règles pour traduire un langage inconnu vers un langage connu. Cela ne donne pas à ces symboles un contenu, une sémantique concrète. En dernier ressort, le seul contenu des théories scientifiques et philosophiques est donné par l'expérience sensible. C'est là le point capital de Kant. Ce qui signifie que la machine, d'après cette deuxième lecture, n'aurait pas moins de raison spéculative que l'être humain.

Je laisse au lecteur le soin de choisir qui, de Kant ou de Platon, il préfère adopter les idées. Dans la vision kantienne, la raison contemplative n'existe pas autrement que comme pure sensation morale et esthétique. Dans la vision platonicienne, elle existe comme une expérience intellectuelle à part entière. Dans un cas comme dans l'autre, on dispose d'un critère intéressant pour distinguer l'esprit humain du fonctionnement de la machine, lié étroitement à l'intuition de l'infini. Je reviendrai sur cette idée plus tard.

3.4 La Machine manque-t-elle d'Entendement ?

Au vu de ce que j'ai avancé jusqu'à présent, il semble recevable de dire de la machine qu'elle a une forme de rationalité, mais pas de structures sensibles. Elle peut en effet manipuler des symboles de façon logico-déductive, et produire des conclusions dans le domaine de la spéculation pure, mais elle ne dispose pas des intuitions de l'espace et du temps. Cela veut-il dire que la machine ne sent rien du tout ?

J'aimerais atténuer cette hypothèse en défendant qu'elle possède tout

juste assez de sensibilité pour produire des jugements synthétiques *a posteriori*, et tout juste assez de raison pour exercer un jugement analytique, mais pas des jugements synthétiques *a priori*. Autrement dit, je voudrais arguer du fait que la machine possède certes une sensibilité et une rationalité, mais pas d'Entendement — entendu comme la faculté de juger du vrai en l'absence d'expériences nombreuses ou de justification logique *a priori*. Cette posture nuancée dont je me fais l'avocat me semble plus avantageuse à plusieurs égards.

Premièrement, elle se montre plus charitable à l'endroit de l'intelligence artificielle. Lorsqu'il s'agit de critiquer un paradigme scientifique, il est habile de concéder à l'adversaire un certain nombre de prémisses préalables et de voir si, en dépit de celles-ci — ou grâce à celles-ci — il est possible de pointer du doigt la présence d'insuffisances. Dans le pire des cas, ces insuffisances apparaissent sous la forme de contradictions entre les prémisses ; mais le plus souvent, il s'agit seulement d'indiquer que le projet initial n'est pas aussi prometteur qu'on le croyait. Cela peut survenir soit parce qu'il s'avère nécessaire d'introduire des prémisses nouvelles, dont certaines sont peu convaincantes, soit parce qu'il y a des angles morts dans la pensée, des zones d'ombre, des naïvetés concernant la façon dont on se représente l'esprit humain, son fonctionnement, ses déterminants essentiels.

Je souhaite concevoir ma thèse de sorte qu'elle soit propice à inspirer le débat, à susciter la polémique. En contestant à la machine la possibilité d'avoir une sensation ou une raison, et en assimilant les algorithmes à des automates sans âmes, similaires à ceux que dépeignait Descartes, qui ne font rien d'autre qu'imiter des raisonnements humains, la discussion est rendue inutile — lorsqu'elle ne revient pas à taper du poing sur la table en affirmant dogmatiquement ses conceptions.

Il est en effet impossible de se mettre à la place de la machine pour savoir si elle "sent" et si elle "pense". Il y a là-dedans quelque chose d'intrinsèquement invérifiable, et ce pour des raisons de définition. La subjectivité, à tout le moins dans notre métaphysique occidentale, est conçue comme un espace strictement privé et individuel. Tout ce qui se passe dans mon esprit ne relève que de moi, et de personne d'autre. Par suite de quoi, il est impossible d'avoir accès à cet espace mental.

Je dis que c'est un fait de notre métaphysique matérialiste, et pas un fait en soi, car il existe des cultures où la croyance en la télépathie, ou la communication avec l'esprit des morts, est chose courante. Il s'ensuit que, pour eux, le problème posé par la subjectivité ne susciterait peut-être pas exactement les mêmes réflexions, les mêmes promesses et les mêmes impasses que pour nous autres^[15]. Même dans ces cultures, il va de soi qu'il existe un

15. On peut notamment penser aux cultures chamaniques, pour lesquelles le rêve permet d'accéder à un espace mental subjectif qui est commun à plusieurs êtres intelligents, aussi bien les animaux, les végétaux que les défunts, et dans lequel la communication est permise. « Toutes les sociétés de tradition orale, en particulier les sociétés à chamanisme, opposent

problème irréductible de communication. Je veux seulement dire qu'il n'est sans doute pas compris de façon aussi radicale.

Le problème de la subjectivité, notons-le bien, existe aussi bien face à une machine que face à un être humain. Je ne peux pas savoir de façon certaine que le lecteur dont les yeux échouent sur ces lignes ressent et pense véritablement quelque chose, ou s'il ne cherche qu'à m'en convaincre par une maîtrise de l'acting à faire pâlir les cabotins. Je pars donc du principe que mon interlocuteur est, tout autant que moi, capable de ces facultés ; et c'est à partir de ce point de départ que peut s'engager un débat proprement philosophique.

Bien sûr, il peut arriver, dans le cours de la conversation, que l'un des interlocuteurs constate, chez son répondant, un manque de sensibilité, de logique ou d'expérience de vie ; en dépit des explications les plus consciencieuses, il peut être difficile de surmonter l'indicibilité, l'incommunicabilité propre à toutes les relations humaines. Chaque fois que ceci se produit, la véritable question consiste à savoir quel est l'élément qui empêche la compréhension mutuelle ? Très souvent, il peut s'agir d'une valeur. L'ensemble du discours humain est généralement conditionné par des représentations morales sous-jacentes, à propos de soi ou du monde. Certaines idées nous sont opaques en raison de ce qu'elles induisent sur la définition de nous-mêmes. D'aucun refuseront d'accepter tel ou tel propos sur la question des hommes et des femmes, car ils ne veulent pas passer, à leurs propres yeux, pour des réactionnaires ; d'autres refuseront d'entendre une critique à l'égard d'un parti politique ou d'une institution qu'ils ont toujours soutenu, car cette critique rendrait caduque (pense-t-ils sans le savoir) le temps qu'ils ont consacré à soutenir cette organisation.

Dans certains cas également, le blocage peut survenir par suite d'un manque d'aptitude. Une lacune de raisonnement logique, de compétence ou d'expérience dans un certain domaine constitue certainement le cas le plus visible. Un expert en hydroélectricité pourra éprouver tant de difficultés à exposer ses idées à l'homme du commun qu'il se verra dans la nécessité, parfois douloureuse, d'éviter certains sujets. Il réalise l'écart existant entre lui et son interlocuteur à la présence de certains indices dans la conversation.

C'est exactement cette stratégie que je souhaite entreprendre avec l'intel-

deux mondes : ce monde-ci, du quotidien, du profane, de l'ordinaire, et le « monde-autre », peuplé de dieux, d'ancêtres, de spectres, de créatures fantastiques et d'êtres de ce monde-ci doués provisoirement de propriétés spéciales, tel le chaman. Le monde-autre est qualifié par chaque société d'un mot que l'on peut rendre par « sacré ». Entre ces deux mondes, on suppose possible une communication et l'on considère que bon nombre d'infortunes - parfois de bonnes fortunes - en sont les conséquences. Les êtres du monde-autre menacent ou agressent les âmes, les corps, les biens ou le milieu. Mais le monde-autre offre aussi aux humains la possibilité de communiquer avec lui et de répondre à ses attaques. Cette communication se fait le plus souvent par le rêve, considéré comme une voix émise et gouvernée par le monde autre qui révèle ainsi la vérité à venir, proférant pronostics et oracles¹⁶.

ligence artificielle. Je veux partir du principe que la machine "sent", qu'elle "pense" et déterminer, par le truchement de certains indices, s'il ne manque pas une certaine aptitude. Cette aptitude n'est pas celle de la sensation pure, ni celle de la raison pure, où la machine imite assez fidèlement l'esprit humain, mais occupe l'exact intermédiaire entre ces deux facultés : il s'agit de l'Entendement, conçu comme une capacité à former des concepts à partir du monde sensible.

Si l'on s'en tient à la théorie kantienne, l'absence de jugements synthétiques *a priori* peut être due à plusieurs facteurs : (a) une lacune au niveau des intuitions pures, que nous avons déjà vues, (b) des lacunes au niveau des formes pures ou (b) l'absence du Je permettant la déduction subjective, c'est-à-dire, l'unité entre toutes les impressions, les sensations, les perceptions qui traversent la conscience. Dans ce chapitre, j'aimerais examiner le point (b). Je réserverai le point (c) pour le chapitre suivant.

3.4.1 Lacunes au niveau des catégories

J'ai déjà discuté la possibilité que la machine n'ait pas d'intuition pure de l'espace et du temps, qui sont les formes pures de la sensibilité, et j'ai montré qu'elle était plausible. J'aimerais donc envisager ici l'alternative d'après laquelle la machine ne soit pas dotée des douze catégories qui sont les formes pures de l'Entendement. Pour rappel, ces dernières se rangent en quatre espèces :

1. **Quantité** : Unité, pluralité, totalité
2. **Modalité** : Possibilité, Existence, Nécessité
3. **Qualité** : Réalité, Négation, Limitation
4. **Relation** : Substance, Causalité et dépendance, Communauté

J'ai suggéré, dans le chapitre 2.4, que ces quatre types de catégories coïncident approximativement avec les quatre ingrédients que François Chollet identifie dans le raisonnement abstrait présidant à la résolution des exercices ARC (les tests de QI de la machine, cf. chapitre) :

1. **Arithmétique** : Capacité à trier les objets en fonction de leur forme, de leur couleur, de leur apparence ou mouvement en s'appuyant sur l'addition, la soustraction et la comparaison.
2. **Persistance de l'objet** : Capacité à repérer des objets et à les manipuler sans qu'un changement d'orientation ou de contexte fasse disparaître l'identification de cet objet.
3. **Géométrie et topologie** : Capacité à réaliser des manipulations spatiales sur les objets, telles que des rotations, des symétries orthogonales, des translations, des déformations, des répétitions, des associations, etc. Il est possible, en particulier, de déterminer la distance entre deux objets.
4. **Orientation vers d'un but** : Capacité à imputer des intentions ou des buts à certains objets.

François Chollet, lorsqu'il identifie ces critères, s'intéresse indirectement au problème qui consiste à identifier ce qui rend le jugement possible au sein des algorithmes. Ce jugement qu'il a à l'esprit est une certaine capacité de généraliser et de penser dans l'abstrait. Il repose sur un certain nombre de briques conceptuelles. Ces dernières ont ceci de particulier qu'elles ne sont pas simplement des lois logiques : elles sont le liant qui réunit, dans une pensée, une forme logique à un contenu sensible. En des termes plus simples, les catégories de l'Entendement sont des formes abstraites qui structurent l'information sensible et n'existent qu'au travers de cette dernière.

La question qui s'impose est celle de savoir si les algorithmes actuels disposent de ce genre de concepts structurels ? Plutôt que de me pencher sur les quatre espèces de catégories à tour de rôle, j'aimerais concentrer mes efforts sur les catégories de la Quantité. Ces dernières sont en effet les plus susceptibles, en raison de leur simplicité, d'être possédées par une machine. Elles consistent à faire une distinction d'espèce entre des jugements généraux et les jugements particuliers tels que ci-dessous :

- P1 : « Ceci est une machine pensante »
- P2 : « Il existe (au moins) une machine pensante »
- P3 : « Toute machine est une machine pensante »

D'un point de vue mathématique, la nuance entre les phrases P2 et P3 peut être obtenue en introduisant les symboles « il existe » et « pour tout », notés respectivement $\exists$ et $\forall$. Ces symboles permettent d'exprimer l'idée qu'une certaine propriété, comme le fait d'être pensant, soit vraie pour une machine M donnée (énoncé particulier) ou pour l'ensemble des machines M envisageables (énoncé général). Muni de ce vocabulaire, les phrases P1, P2 et P3 se réécrivent :

- (i) M est une machine pensante (toute machine pense)
- (ii) $\forall M : M$ est une machine pensante (toute machine pense)
- (iii) $\exists M : M$ est une machine pensante (il existe une machine pensante)

Il est possible de donner à cette expression une forme plus aboutie, plus précise, disons, en utilisant la notation $p(M)$ pour exprimer : « La machine M pense¹⁷ ». Cela donne :

- (i) $\forall M p(M)$ (toute machine pense)
- (ii) $\exists p(M)$ (il existe une machine pensante)

17. D'un point de vue tout à fait rigoureux, le symbole p est ce qu'on appelle un prédicat. C'est une fonction qui prend une variable (ici une machine M) et lui associe soit la valeur *vraie* soit la valeur *fausse*.

Ces symboles doivent être utilisés dans un langage suffisamment précis pour être compris par une machine. Un exemple de langage formalisé est la logique des prédicats. Au sein de cette dernière, on observera quelques propriétés remarquables. À titre d'exemple, « si toutes les machines pensent », alors « il est faux qu'une machine ne pense pas. ». Inversement, « s'il existe une machine qui pense », alors « il est faux que toutes les machines ne pensent pas ». Ces deux lois que je viens de citer peuvent s'écrire de la manière suivante :

- (i) $p(M) \rightarrow \neg \exists \neg p(M)$
- (ii) $\forall M p(M) \rightarrow \neg \exists \neg p(M)$
- (iii) $\exists p(M) \rightarrow \neg \forall \neg p(M)$

Je me dois de préciser, à l'intention des non logiciens, que l'opérateur $\rightarrow$ est ce qu'on appelle l'implication logique : si la phrase à gauche de la flèche est vraie, alors celle à droite est vraie également. Par ailleurs, le symbole $\neg$ exprime la négation. Lorsqu'il est placé devant la phrase, il veut dire que la phrase est fausse. Lorsqu'il est placé devant $p(M)$, il signifie que la machine M ne pense pas. Pour être tout à fait clair, $p(M)$ doit donc se lire : « la machine M pense » tandis que $\neg p(M)$ se lit : « La machine M ne pense pas. ».

Il ne s'agit que d'un petit exemple des lois logiques qu'il est possible de construire avec les quantificateurs $\forall$ et $\exists$. En recherchant une à une toutes les autres lois de la pensée de la logique des prédicats, et en les formalisant, on obtient un système logique compréhensible par une machine. Ce système permet de distinguer les énoncés généraux des énoncés particuliers.

3.4.2 Logique générale vs transcendantale

La question résiduelle, et philosophiquement la plus significative, consiste à se demander si la machine obtenue à l'aide de toutes ces lois logiques est, oui ou non, munie de catégories *a priori* ? Kant faisait en effet une distinction entre la logique générale, qui consiste à envisager objectivement les lois du raisonnement, et la logique transcendantale, qui s'intéresse à la façon dont ces lois peuvent s'appliquer aux contenus sensibles de la connaissance au sein d'un sujet pensant. La question posée par l'intelligence artificielle, dans un jargon qui vous permettra de briller dans vos conversations mondaines, consiste donc à se demander : quels sont les ingrédients nécessaires à la logique transcendantale ?

En guise de premier point de réflexion, on peut exciper du fait que l'être humain, contrairement à la machine, utilise des énoncés généraux et particuliers sans en connaître les règles axiomatiques. Il peut même se montrer habile en l'absence de toute connaissance des symboles $\forall$ et $\exists$. C'est ainsi que certains esprits intuitifs, n'ayant pas de formation mathématique, se

débrouillent prodigieusement bien en logique. Si donc la machine a besoin qu'on lui explique tout ce qu'il est permis de faire ; de notre côté, nous faisons de la logique tout comme Monsieur Jourdain faisait de la prose : sans s'en rendre compte !

C'est la raison pour laquelle, d'un point de vue kantien, ces catégories sont, chez l'être humain, des structures *a priori* de la pensée. Elles existent lorsque l'on pense, même si on ne les connaît pas. Elles se diffusent pour ainsi dire dans chacun de nos contenus de pensée. J'aimerais illustrer, à l'aide d'un exemple éloquent, ce fait que la pensée humaine est sculptée par la modalité. Si je vous dis : « Le matin, le soleil se lève » ou « Le matin, l'armée américaine débarque sur les côtes normandes », vous allez spontanément associer ces phrases à l'un des deux quantificateurs $\forall$ ou $\exists$. Vous traduirez dans votre tête « l'armée a débarqué en ce matin du 6 juin 1944 » et non pas « tous les matins, l'armée débarque ». En revanche, vous penserez sans doute « Le soleil se lève tous les matins » et non « Le Soleil s'est levé le 6 juin 1944 ! »

Bien entendu, l'esprit humain est sujet à l'erreur. C'est pourquoi il peut arriver que nous n'assignons pas toujours la même modalité que notre locuteur à la phrase que nous entendons. Il peut également se produire que nous ne fassions pas les bonnes déductions, que nous trahissions, pour ainsi dire, les lois de la logique. C'est ce à quoi l'on arrive si, porté par les apparences immédiates, nous nous confondons dans le raisonnement suivant :

1. Le soleil s'est levé le 6 juin 1944.
2. L'armée américaine a débarqué le 6 juin 1944.
3. Donc l'armée débarquera lorsque le soleil se lèvera.

Cette corrélation entre deux faits contingents, le lever du soleil et la venue de l'armée, n'est (heureusement) pas une conséquence logique des phrases d'après lesquelles « l'armée peut débarquer un matin » et « le soleil doit se lever tous les matins ». Il arrive pourtant régulièrement, dans des cas certes plus complexes, que l'être humain commette cette méprise. Un exemple éloquent est cette étude qui démontre que les pays où l'on mange davantage de chocolat sont plus susceptibles de donner naissance à de futurs prix Nobel. Vous ne pouvez pas en inférer que le chocolat rend plus intelligent, sans quoi la Belgique devrait détenir le palmarès !

En dépit de nos erreurs fréquentes, il existe néanmoins un consensus à propos de ce qui doit être tenu pour un raisonnement valide ou non. Lorsque l'on met plusieurs êtres humains autour de la table, et que l'on attend suffisamment longtemps, et exception faite des différends de nature idéologique ou personnelle (qui représentent parfois l'essentiel du conflit), ils finissent généralement par converger vers un commun accord¹⁸. Il peut

18. Cette position a reçu une expression plus complète de la part de Habermas, qui

certaines arriver que ce processus prenne plusieurs générations — les théories scientifiques sont les ruines de ces guerres d'interprétations — mais force est d'admettre la présence d'une certaine robustesse, ou d'une stabilité, dans les conclusions.

Ces deux caractéristiques que j'ai énoncées, celle d'après laquelle (en caricaturant un peu) (1) l'erreur est humaine, mais (2) la vérité triomphe toujours, ne semblent pas être l'apanage des machines (tout du moins pas des machines les plus rigides ; je discuterai plus loin le cas des réseaux de neurones avancés tels que ChatGPT). Tel algorithme programmé pour réfléchir avec les modalités de la « possibilité » et de la « nécessité » ne se trompera pas eu égard aux lois de la logique, mais il ne s'accordera jamais avec un algorithme programmé pour réfléchir avec une logique paraconsistante, admettant qu'une chose puisse à la fois être vraie et fausse. À quoi ressemblerait un débat entre des êtres faits de câbles et d'engrenages, qui restent fermement accoudés à leurs principes logiques ? Il semblerait que la flexibilité de l'être humain soit indissociable de sa capacité à faire des erreurs. Sa pensée est certes munie de structures logiques, mais celles-ci ne sont pas strictes. Ce ne sont que des façons de comprendre le monde.

En conclusion, doter l'ordinateur des catégories *a priori* de l'Entendement ne revient pas, purement et simplement, à le munir des règles permettant la manipulation des symboles. Une telle approche se limiterait à définir une logique générale, et non ce que Kant appelait une logique transcendantale. C'est-à-dire qu'elle s'intéresse à la validité en soi des raisonnements, mais non à la façon dont les données elles-mêmes sont mises en forme par ces lois. Or c'est uniquement ce deuxième aspect, celui du rapport de la logique avec le monde réel, qui est proprement en jeu dans les catégories *a priori*. Il semblerait également que ce soit cette capacité de généralisation intuitive, qui rapporte des concepts logiques fondamentaux à des ensembles d'informations brutes, que cherchent à mesurer les tests de QI ou les exercices ARC de François Chollet, avec des résultats plus ou moins probants.

Je tâcherai de montrer, plus loin dans cette thèse, que le paradigme actuel de l'informatique manque ce but : tantôt il donne à la machine des lois formelles à digérer analytiquement comme une recette de cuisine, tantôt il crée des corrélations convaincantes, mais *a posteriori*. Ces deux voies sont celles (a) des systèmes experts, qui pensent à partir de lois logiques ; et (b) des réseaux de neurones, qui établissent des associations statistiques. Je tâcherai de montrer que la voie du milieu, celle de la synthèse *a priori*, est manquée dans chacune de ces deux démarches.

s'intéresse aux conditions de la rationalité communicationnelle, ou encore, à l'ensemble des critères permettant de dire d'un débat qu'il aboutira à un consensus. Cela exige notamment deux ingrédients : un principe de débat assurant une forme de représentativité et un principe universel assurant une forme d'objectivité et d'ouverture aux intérêts de tous.

3.5 La Machine est-elle un Je ?

Qui n'a jamais eu l'étrange ressenti, en utilisant ChatGPT, que les réponses affichées sur le terminal, quoi qu'impressionnantes, manquaient de personnalité ? Tout se passe comme si la machine synthétisait des expériences, des opinions, des réflexions qui jonchaient la toile infinie de l'Internet, mais sans jamais rien apporter de première main. Ses affirmations sont certes structurées et bien informées, elles sont nuancées, drôles à l'occasion, mais elles restent à la surface de la pensée. On peut avoir le sentiment que l'algorithme nous offrait une image floutée de la réalité, un gros plan, une esquisse statistique, qui manquait de contours, de vie, de caractère. À l'instar d'une onde quantique, ses productions brassent des centaines de possibilités, elles oscillent entre différents points de vue, mais sans jamais prendre une situation précise dans le monde. Pour paraphraser le titre du livre de Robert Musil, ChatGPT est un homme sans qualité ; un homme de pures possibilités.

C'est pourquoi, après avoir fait l'analyse des catégories de l'Entendement, il me reste à demander si les machines que nous connaissons aujourd'hui peuvent avoir un Je. Le Je, chez Kant, est défini comme le fait même que, derrière chaque acte de conscience isolé, c'est bien d'une même conscience qu'il s'agit. En l'absence de Je, en effet, et c'est peut-être ce qui se produit chez certains sujets victimes de schizophrénie, les contenus de la pensée, de la sensation, de l'émotivité seraient vécus comme des expériences isolées, déconnectées les unes des autres, et ne nous appartenant pas en propre. Il serait dès lors impossible, lorsque l'on est placé face à un paysage, et que nos yeux errent d'un élément à l'autre de celui-ci, d'identifier la forme d'une montagne se découpant dans le brouillard, ou d'un navire perdu à l'horizon. Ce procédé est impossible tout simplement parce que ce n'est plus la même conscience qui regarde la montagne et qui regarde le bateau. Il n'y a pas de liaison possible entre ces deux contenus.

Plus fondamentalement encore, l'absence de Je empêche de voir un bateau ou une montagne, car ces objets eux-mêmes sont composés d'une diversité d'impressions fugaces : des points de couleurs, des formes, qu'il est impossible de relier. Une conscience sans le Je serait une parfaite illustration de la création continue. À chaque instant, Je serais, moi, un autre Je ; si bien qu'il n'y aurait aucune mémoire, ni aucun concept. Le monde, si tant est qu'il existerait encore en nous — on peut douter qu'une telle conscience soit conscience de quelque chose — s'apparenterait dans le meilleur des cas à un grand kaléidoscope.

Dans la dernière partie de cette thèse, j'entreprendrai de donner une définition phénoménologique du Je, basée sur cette lecture kantienne ; et je tenterai de montrer qu'il repose sur un principe de différenciation qui exige une forme intrinsèque d'ambiguïté. Il y a en effet dans la conception du monde une capacité de l'esprit à comparer la grande fresque des sensations

qui échoient à la conscience de façon à identifier, en son sein, des vastes ensembles d'objets. Le bateau, la montagne nous apparaissent justement parce qu'ils sont différenciés de l'arrière-plan sur lequel ils nous apparaissent premièrement. Cela implique, d'un point de vue fondamental, que la perception n'est pas donnée comme un ensemble de points isolés que notre esprit relie pour former des catégories. Au contraire, la perception se présente d'abord comme un Tout, après quoi seulement nous la divisons en objets divers et variés. C'est pourquoi nous éprouvons une sorte de vertige, de fascination, lorsque nous sommes premièrement plongés dans l'observation d'un paysage. Il semble dépasser le contenu de notre pensée. Il nous touche, nous transporte. Et ce n'est qu'après ce bref moment de perdition, qui est proprement le sens du beau, que nous entreprenons tout aussitôt, inconsciemment, de décomposer cette unité indicible en tout autant d'objets connus.

Dans la machine, au contraire, ce vertige ne saurait que difficilement avoir lieu. La raison en est qu'on lui offre l'information sensible sous une forme prédigérée. La structure du langage informatique impose de dire à l'algorithme : « voici une image : il s'agit d'une matrice de points. Chaque pixel est représenté par trois nombres entiers définissant la couleur, un nombre entier définissant la luminosité, un autre définissant l'albédo, etc. » Ce faisant, on a déjà scindé l'image à analyser en un ensemble d'unités minimales. Ce n'est plus un Tout qui se donne à la conscience, mais des contenus isolés les uns des autres.

3.6 La Machine pense-t-elle différemment ?

Avant de terminer cette partie critique sur les limitations de la machine, il me faut encore toucher un mot sur la dernière alternative que j'ai soulevée à propos du jugement des machines. Celle-ci, de loin la plus favorable d'entre toutes au projet de l'intelligence artificielle, consiste à admettre que les algorithmes aient la faculté de juger, aussi bien en ce qui concerne les jugements analytiques, synthétiques *a posteriori* que les jugements synthétiques *a priori*. Sous cette hypothèse, néanmoins, on ne peut pas nier l'observation de différences dans les modes de raisonnement entre l'humain et la machine.

D'une part, la machine sait faire nombre de choses dont l'humain se montre incapable : elle a, entre autres faits incontestables, une puissance de calcul et une exactitude contre lesquelles il est difficile de rivaliser. Mais d'autre part aussi, la machine échoue à identifier certaines catégories, ou à procéder à certaines abstractions, qui sont pourtant naturelles pour nos homologues humains. C'est ce qui se produit lorsqu'elle sous-performe aux exercices ARC de François Chollet ou, cas plus célèbre encore, lorsqu'elle identifie des personnes noires comme étant des « gorilles¹⁹ »

19. Ce fait divers, survenu le 1er juillet 2015, a défrayé la chronique, lorsque Jacky Alciné et son amie noire ont constaté que l'algorithme les cataloguait comme des "gorilles". On

Si donc on accepte que la machine soit capable des mêmes types de jugement que ceux dont l'être humain se montre capable, force est de constater qu'elle ne les exerce pas nécessairement de la même façon. Des différences existent quant à leur application, qui conduisent à tout autant de disparités comportementales observables. Quelle est exactement l'origine de ces différences ? Comment peut-on les cerner ?

Un premier ordre de différence a déjà été évoqué : c'est lui qui existe, du point de vue de l'Entendement, à propos des lois de la logique. J'ai montré que différents formalismes pouvaient être utilisés pour programmer une machine, conduisant à des lois bien étrangères les unes aux autres. Certaines de ces lois répondent aux formes *a priori* de l'esprit humain reconnues par Kant : c'est le cas de la logique classique. Mais d'autres lois peuvent y contrevenir, comme lorsqu'on dit qu'une chose peut à la fois être et ne pas être. Ce genre de proposition choque généralement l'intuition humaine, encore qu'on puisse dans bien des cas lui trouver un sens.

Dire de la machine qu'elle est capable de jugements synthétiques *a priori* ne veut pas dire autre chose que ceci : les données sensibles sont mises en forme de façon *a priori*. Il existe une structure, un ordre, un prisme, qui s'interpose entre la sensation et l'intelligence. Si, à nous autres êtres humains, les points de l'espace nous apparaissent sous formes de carrés, de cercles, de triangles, qu'en serait-il aux yeux d'une machine ? Ne pourraient-elles pas détenir des formes qui lui sont propres ?

Le test ARC, sous cette hypothèse, ne serait pas tant un test d'intelligence qu'un test de discrimination, analogue à celui utilisé dans la nouvelle de Philippe K. Dick intitulée : « Les androïdes rêvent-ils de moutons électriques²⁰ ? » Ils permettraient d'identifier, en exhibant des problèmes significatifs, si l'on est face à un être humain ou à une machine habilement dissimulée.

Il y a quelque chose d'arbitraire dans les matrices de Raven utilisées par les tests ARC. Bien que la réponse attendue nous paraisse intuitive, le lien liant les inputs aux outputs ne relève pas d'une pure déduction logique, mais d'une simple induction. Or l'induction dépend en partie de nos habitudes de pensée. Nous sommes habitués à nous représenter des formes géométriques telles que des carrés ou des triangles et à leur faire faire des transformations par symétrie. Nous y sommes habitués, car il s'agit de notre façon instinctive de structurer les informations. En revanche, nous sommes moins aptes à

aurait pu penser que le problème avait été corrigé. Or en 2023, l'application ne permet toujours pas d'identifier les singes. On verra qu'avec la programmation des réseaux de neurones, il est en effet difficile de fournir des garanties. Même lorsque le système présente de bons résultats sur sa base de données de test, il n'est pas exclu qu'une erreur soit commise en présence d'un cas réel.

20. Dans cette nouvelle, qui a donné naissance à Blade Runner, les robots sont indistinguables des êtres humains, si bien que certaines personnes sont chargées de les identifier à l'aide d'un dispositif expérimental.

réaliser des calculs complexes. C'est pourquoi nous réussissons facilement les tests ARC, mais que nous peinerons probablement à compléter la série de nombres suivante :

3958585969, 11875757907, 35627273721, 106881821163...

Réfléchissez quelques instants sur ce problème avant de voir la solution. Il est peu probable, je crois, que vous trouviez 320645463489. Il ne s'agit pourtant que de multiplier le nombre initial par trois à chaque nouvelle étape. Cette opération est complexe pour un être humain, mais pas pour une machine. Il est raisonnable de penser qu'un ordinateur puisse obtenir de meilleurs résultats à cet exercice que l'être humain. Il serait possible, en théorie, de faire une batterie de tests, du type matrices de Raven, telle que seuls les ordinateurs obtiennent de bons résultats. Ce serait un test visant à déterminer si le sujet a la même intuition que la machine.

Ces remarques que je fais n'invalident pas la grande idée de ma thèse : à savoir que la machine ne possède pas la même façon de structurer l'information. Elles représentent en revanche une objection à la simplification, philosophiquement naïve, qui consiste à voir dans les tests ARC un test d'intelligence général. Cette simplification du reste, existe déjà largement dans l'univers controversé de la psychométrie. Des auteurs sont connus pour avoir fait des affirmations optimistes à propos de la pertinence du test de QI, et des matrices de Raven, tout particulièrement pour mesurer l'intelligence des individus. On remarque cependant que le QI moyen est très variable en fonction des pays et des cultures. Certaines régions d'Afrique noire présentent des statistiques avoisinant les 60 points de QI, ce qui correspond au seuil de débilité mentale. Doit-on en conclure que les personnes ressortissant de cette ethnicité sont des êtres humains attardés — avec toute la connotation péjorative accompagnant l'usage de ce terme ? Ou faut-il admettre plutôt que le QI dépend de la culture et de la façon dont les individus envisagent certains problèmes²¹ ? Des tests *culture fair* ont vu le jour ces dernières années, pour restreindre au maximum, dans la formulation et le choix des questions, les éléments concurrents à améliorer les performances des personnes originaires des pays du Nord. Il est légitime de se demander si les tests ARC ne contiendraient pas des biais, dont certains sont en faveur de l'être humain et d'autres en faveur de la machine. J'en évoque ici quelques-uns :

- **Formes géométriques** : Les exercices semblent reposer sur une culture géométrique qui est favorable à l'être humain, et en particulier ceux issus des pays du Nord. Les formes à manipuler s'appa-

21. En 1948, Raymon Cattell construisit un test de QI sans tâche verbale, qui était censé surmonter les biais culturels. Dès les années 70, des études soulignaient toutefois la présence de biais (voir notamment JOHN F. IRELAND, MARVIN W. KAHN, *How Fair Is the Culture I.Q. Test ?*, Sage Journal, V.25-1, April 1979).

rentent aux figures de la géométrie euclidienne : carrés, triangles, etc. Il pourrait être intéressant, à titre purement spéculatif, de se demander si la machine présente des prédilections lorsque les formes sont présentées dans un certain type de géométrie ?

- **Temps de réponse :** L'absence de limite de temps, ou une limite de temps élevée, est favorable aux êtres humains. Par ailleurs, cela encourage une réponse pondérée, plutôt qu'une réponse instinctive. En baissant le temps alloué, il est plausible que l'on ne mesure plus tellement le raisonnement en tant que tel que la perception immédiate. Pour discriminer l'être humain de la machine sur la perception du temps, notons que l'on pourrait envisager un test de simultanéité. L'objectif de ce test serait de déterminer si deux événements sont, oui ou non, simultanés. Supposons que l'on soit confronté à un son (un bruit de vase brisé, une porte qui grince, un coup de feu, un coup de tonnerre...) et à une vidéo (un vase qui chute, une porte qui s'ouvre, une balle qui touche une cible, un éclair...). Les sons et les images sont présentées séparément. Le test revient à demander au sujet de déterminer quand le son est, le plus probablement, censé survenir dans la vidéo. Par exemple, le bruit de vaisselle brisée doit s'entendre au moment où le vase touche le sol, tandis que le bruit de l'éclair doit surgir quelques secondes après le flash lumineux. L'esprit humain est habitué à synchroniser les stimuli visuels et auditifs. En revanche, ce ne serait pas un test facile pour la machine, car il lui faut comprendre ce qui est représenté dans la vidéo.
- **Pixels :** Les exercices sont présentés sous forme de pixels, ce qui est avantageux pour la machine. On pourrait se demander ce que deviendraient les performances si les figures à analyser étaient des formes continues ? Dans ce cas, il faudrait affiner le protocole pour déterminer, avec le moins d'ambiguïté possible, quelles sont les réponses admissibles. Cela peut être obtenu soit en construisant des réseaux de neurones chargés de définir les réponses possibles, soit en se référant à un jury humain, soit en hybridant les deux approches.

Il est possible de mobiliser la théorie du jugement de Kant pour justifier une différence de fonctionnement entre l'esprit humain et les algorithmes :

- Une *machine* encode l'information sous forme d'unités insécables (pixels, bits, etc.) ; elle cherche à associer statistiquement ces unités pour former des concepts.
- *L'être humain*, au contraire, appréhende l'information comme un tout ; il utilise des concepts pour isoler des objets au sein de ce tout.

Tout se passe donc comme si l'esprit humain disposait de structures conceptuelles qui modifiaient sa perception globale, en connectant les parties au tout, là où les algorithmes, n'ayant pas de concepts préalables, se contentaient d'informations objectives et indépendantes les unes des autres.

J'ai illustré ces différences de fonctionnement au moyen de plusieurs illusions perceptives (Kanizsa, Ponzo). Dans la prochaine partie, il s'agira de se demander à quoi ressemblent précisément les structures conceptuelles préalables de l'être humain ? J'en profiterai pour critiquer la théorie de Kant.

Révision du concept de jugement

Vers une théorie Méréologique

Chapitre 4

L'Oubli de l'Entendement : Mort du jugement kantien

Arrivé à ce point de notre cheminement intellectuel, peut-être est-il judicieux de récapituler les mérites de la critique kantienne vis-à-vis de la question de l'intelligence artificielle ? Je pense avoir suggéré, en évitant les technicités superflues, qu'il existe en philosophie trois espèces de jugement, reposant chacune sur trois facultés de l'esprit humain : les jugements synthétiques *a posteriori* relèvent de la Sensibilité, les jugements synthétiques *a priori* ressortent majoritairement de l'Entendement et les jugements analytiques reposent sur la Raison. Je dis que les jugements synthétiques *a posteriori* relèvent « majoritairement » de l'Entendement, et non « strictement », parce qu'ils sont à cheval entre la Sensibilité et l'Entendement. En fait, tous les jugements synthétiques, qu'ils soient *a priori* ou *a posteriori*, utilisent les données du monde sensible, mais de façon différente : les jugements *a posteriori* utilisent l'expérience pour parvenir à leur conclusion, tandis que les jugements *a priori* se fondent sur les catégories de l'Entendement. Ils ne naissent certes pas avant que l'être humain ait fait l'expérience des nombres, des triangles et des angles, mais leur vérité ne dépend pas de cette expérience.

Pour exemplifier concrètement la définition de ces trois types de jugement, on peut dire que les jugements synthétiques *a posteriori* sont des faits empiriques, que seuls nous révèlent l'expérience quotidienne d'un monde adipeux, incongru, râpeux : des propositions telles que « cette table est blanche » ou « le bronze est un alliage de cuivre et d'étain », ou encore « la production de CO_2 provoque le réchauffement climatique » sont obtenues en observant des concordances, des récurrences dans l'ordre des phénomènes. Il n'y a rien dans le concept de « CO_2 » qui nous permette de déduire, par une nécessité irréfragable, qu'il y a réchauffement. Si en effet cette vérité découlait uniquement des concepts, elle aurait moins longtemps soulevé le débat. Il se trouve hélas qu'il a fallu faire des mesures, poser des observa-

tions, tester des hypothèses, tant et si bien que l'incertitude restait possible ; c'est le propre des faits contingents que d'être susceptibles d'être révoqués en doute.

Les jugements analytiques, quant à eux, n'expriment rien d'autre que ce qui est déjà contenu dans les termes du discours. Si je dis par exemple que « le cheval blanc de Napoléon est blanc », il n'y a aucune information nouvelle. Il en va de même avec les syllogismes : si je sais que tous les carrés sont des rectangles et que les rectangles sont des quadrilatères, je peux en déduire que les carrés sont des quadrilatères. Ce faisant, je n'apprends rien sur le carré que je ne sache pas déjà. Un raisonnement déductif, aussi savant soit-il, ne fait jamais qu'éclaircir des vérités qui étaient déjà contenues, en puissance, dans les axiomes de la théorie. Il nous place dans la position de l'esclave du Ménon, qui découvre, en suivant le jeu de questions adressées par Socrate, des conclusions logiques. « Ces opinions », commente Socrate, « viennent de se réveiller en lui comme un songe. Et si on l'interroge souvent et de diverses façons sur les mêmes objets, sais-tu bien qu'à la fin il en aura une connaissance aussi exacte que qui que ce soit¹? »

Les jugements synthétiques *a priori*, pour finir, sont des raisonnements qui ne se réduisent ni aux lois de la logique, ni aux résultats de l'expérience. Ils trahissent l'existence de structures fondamentales dans l'esprit humain. À titre d'exemple, Kant évoque la proposition d'après laquelle « $5 + 7 = 12$ » ou du fait que « la somme des angles d'un triangle est égale à 180 degrés ». Ces relations, d'après lui, ne sont pas des vérités analytiques. Il n'y a rien dans les concepts de « triangles » et « d'angles » qui impliquent, de toute nécessité, cette relation. Ces vérités mathématiques ne sont pas davantage des propriétés imputables aux objets du monde. Elles découlent plutôt des outils conceptuels que nous manipulons pour comprendre le monde : à savoir l'intuition du temps, l'intuition de l'espace et les catégories de l'Entendement. Ces outils ne sont hélas pas visibles directement. Ce serait trop facile. Comment voir le prisme au travers duquel le monde se donne à voir ? Si vous portez des lunettes déformantes depuis votre naissance, vous ne pouvez pas savoir que le monde est déformé ni que vous portez des lunettes. En revanche, vous observerez des constantes, des lois optiques, entre toutes les choses qui vous tombent sous les yeux. Les objets présenteront des courbes, des angles, des points focaux, des symétries caractéristiques. Ce n'est qu'après coup que vous pouvez soupçonner que ces lois ne sont pas propres au monde lui-même, mais à la structure de vos lunettes. Cela veut dire que l'être humain doit être mis au contact du monde avant de pouvoir envisager des propriétés mathématiques à propos des angles, des triangles, des nombres. Mais dès lors qu'il découvre ces vérités géométriques et arithmétiques, il a l'impression qu'elles ne dépendent pas de l'expérience du monde. Ce sont les conditions *a priori* de toute expérience ; ce sont des lois inhérentes aux structures de

1. PLATON, *Ménon*, trad. Cousin, Philosophie, Septembre 2009, p. 65.

la pensée humaine.

J'ai défendu l'idée que c'est l'absence de jugement synthétique *a priori*, et non l'absence de jugements synthétiques *a posteriori* ou de jugements *analytiques*, qui serait à l'origine des différences de fonctionnement entre l'esprit humain et la machine. L'analyse quelque peu technique que j'ai réalisée des exercices ARC de François Chollet semble compatible avec l'idée que l'intelligence artificielle, quant à ses paradigmes actuels, n'est pas en mesure de structurer l'information sensible de façon intuitive et catégorique. Qu'est-ce que j'entends par là ? Que le prisme conceptuel de l'esprit humain n'a rien de comparable au prisme conceptuel de la machine. Lorsque nous programmons un algorithme, nous faisons l'erreur de penser que les structures spatiales et temporelles du monde, telles que les carrés, les cercles ou les instants, existent dans le monde tel que nous les voyons. Nous souhaitons donc créer une machine qui soit capable de les identifier. Or, si l'on prend au sérieux la grande idée de Kant, il est possible que ces formes ne découlent pas uniquement de la juste analyse du monde, mais aussi de la façon dont l'information est structurée au sein de notre esprit.

Force est de constater que la structure des données, chez l'être humain et chez la machine, présente des différences substantielles. Au sein des algorithmes, les données sont fournies sous forme d'unités discrètes indécomposables : les pixels de l'image, les intervalles de temps d'un signal audio, etc. Ces unités sont traitées comme s'il s'agissait nécessairement des fondements minimum de la perception. Elles sont dotées de propriétés intrinsèques (couleur, intensité sonore...) qui ne sont pas influencées par les autres unités du voisinage. Le problème qui se présente au programmeur est de comprendre comment ces points peuvent être traités par un algorithme de façon à identifier les relations existant entre eux : sont-ils disposés de façon à former un carré ? trahissent-ils une profondeur spatiale ? Le problème à un million de dollars de l'intelligence artificielle, pour le dire sous une forme lapidaire, consiste à comprendre comment il est possible de reconstruire une vision d'ensemble². Tout se passe en effet comme si la machine ne possédait que des morceaux de perception et devait les recomposer pour obtenir une perception complète. Pour le dire à l'aide d'une image, la structure de la perception artificielle est similaire à un tamis. Toute onde lumineuse qui passe au travers de ce tamis ressemble à des corpuscules échoués au hasard sur une page blanche (voir figure 4.1).

Chez l'être humain, en revanche, il se pourrait que la perception ne soit pas composée d'unités indivisibles. Le découpage du réel n'est pas comparable au traitement digital, qui extrait l'information d'une image, d'un son, d'une perception, sous forme discrète, en la découpant en petits intervalles

2. L'ensemble des prix décernés par François Chollet et ses associés en cas de succès aux tests ARC s'élève en effet à 1,1 million de dollars. Ces tests, comme on l'a vu, demandent de généraliser des structures à partir d'un très faible nombre de données (voir <https://arcprize.org/research>).

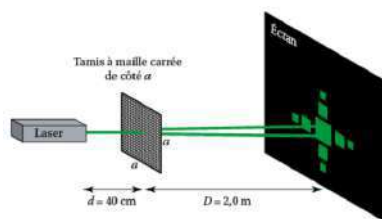


FIGURE 4.1 – La structure *a priori* de la perception des machines peut être comparée à un tamis. Toute information est discrétisée et représentée sous forme d'une grille de points indépendants les uns des autres.

de temps, de distance, et en assignant des valeurs numériques à chacun de ces tronçons. Pour l'être humain, l'information est structurée de façon plus holistique et plus processuelle. L'esprit est attentif aux contrastes existant au sein de la perception d'une image ou d'un son. Il découpe le tout de la perception en utilisant des outils familiers (objets, causalité, figures géométriques...). C'est donc parce que nous avons, en nous, la catégorie du triangle, que les données vont prendre à nos yeux la forme d'un triangle ; et c'est ultérieurement, par le biais de l'analyse minutieuse, que ce triangle va nous apparaître comme un simple nuage de points. Nous voyons d'abord des formes clefs, après quoi seulement nous observons des imperfections. Du reste, il ne semble pas y avoir de limite assignable au processus de décomposition. Nous pouvons diviser une perception en isolant des différences, des contrastes en son sein, puis revenir à la perception de l'ensemble. Autrement dit, je veux poser l'hypothèse qu'il existe un mouvement de va et vient, dans la perception de l'Homme, qui va du tout aux parties, et qui remonte des parties vers le tout.

Ce double processus descendant et ascendant est rendu possible, si l'on en croit la théorie kantienne, par l'intuition pure de l'espace, l'intuition pure du temps et les catégories de l'Entendement. Ces formes sont, dans une certaine mesure, vérifiables scientifiquement — ou tout au moins compatibles avec certains résultats scientifiques. En neuroscience, la théorie de Marburg tente d'identifier, dans l'expérience humaine, l'existence de structures qui déterminent la perception *a priori*. Cette ligne de pensée trouve ses origines chez Cassirer^[3]. Il a donné naissance, entre autres, à la théorie du « Cerveau Kantien^[4] » :

Le modèle du « cerveau kantien » fusionne la philosophie d'Emmanuel Kant avec les recherches modernes sur le cerveau, en proposant que celui-ci génère activement des prédictions et des

3. KATHERINA KINZEL, *Neo-Kantian conceptualism : between scientific experience and everyday perception*, British Journal for the History of Philosophy, 08 Aug 2023.

4. VALÉRY YEVAROUSKY, *The "Kantian Brain" Concept : Modeling the Limit of Human Perception and Existence*, septembre 2023.

interprétations basées sur des connaissances préalables. Il met l'accent sur l'imagination, la créativité et la spontanéité dans la perception et la cognition.

Il n'est toutefois pas certain que la manière dont ces formes sont définies par Kant épouse de façon adéquate le progrès moderne de la science. L'intuition pure de l'espace, notamment, se range dans une vision euclidienne du monde, alors que la physique travaille désormais avec des géométries non classiques, où la plus courte distance entre deux points n'est pas la ligne droite. Par ailleurs, les catégories de l'Entendement véhiculent une vision universaliste de l'esprit humain, où chaque individu naît doté des mêmes structures logiques. Or des travaux en anthropologie semblent indiquer qu'il existe des cultures ne connaissant pas le concept de nombre. Ils utilisent uniquement les catégories « un », « deux » et « plusieurs ». Plus radicalement encore, il y a des traces en Inde de systèmes logiques assez sophistiqués admettant qu'une chose puisse à la fois être et n'être pas. Il ne faut d'ailleurs pas se rendre de côté du Gange pour en faire la découverte. Dans la Grèce hellénistique, les Mégariques disputaient déjà à Aristote le principe du tiers exclu, prétendant que soit une chose est vraie, soit elle est fausse, mais qu'il n'existe pas de milieu⁵.

C'est pourquoi il est peut-être judicieux (1) de critiquer les formes *a priori* de Kant et (2) de les redéfinir de telle sorte que leur expression soit plus compatible avec l'évolution de la science moderne — et de l'informatique en particulier. À cette fin, il s'agit tout d'abord de comprendre pourquoi, en dépit de la notoriété dont la théorie kantienne a joui pendant plus d'un siècle, les jugements synthétiques *a priori*, et plus particulièrement la faculté de l'Entendement, semblent avoir peu à peu déserté la conception scientifique contemporaine du monde ? On peut également se demander en quoi cette transformation historique progressive a donné naissance aux structures épistémologiques utilisées par les algorithmes actuels ?

S'il existait certainement encore des néokantiens au début du vingtième siècle, lorsqu'il s'agissait de s'interroger avec Niels Bohr des interprétations de la physique quantique⁶; si la notion de facultés *a priori* reste très opérationnelle encore aujourd'hui dans les neurosciences, au moment d'identifier, par exemple, les aires cérébrales relatives à l'organisation spatiale des représentations humaines; ces réminiscences de l'Entendement ne constituent que des phénomènes anecdotiques. Force est de constater que l'existence de formes pures et d'universaux fait peur à notre relativisme frileux, plus soucieux de comprendre en quoi l'Homme (et ses observations) sont la mesure de telle ou telle chose que d'identifier des constantes universelles.

5. Sur cette matière, consulter l'excellent ouvrage de LUKASIEWICZ, *Du principe de contradiction chez Aristote*, 1910, trad. fr. de Dorota Sikora, préface de Roger Pouivet, Paris, L'Éclat, 2000

6. Se référer notamment au chapitre intitulé *Bohr's Complementarity and Kant's Epistemology*, issu de l'ouvrage *Niels Bohr, 1913-2013*.

J'aimerais défendre dans cette section qu'un des facteurs les plus importants ayant conduit à « l'oubli de la faculté de l'Entendement », que j'appellerai plus simplement « l'oubli du Beau », est à rechercher dans la structure même des mathématiques. Il se trouve en effet que les deux exemples de prédilection de Kant, pour illustrer l'existence de jugements synthétiques *a priori*, s'inspirent de l'arithmétique et de la géométrie. Kant écrit à plusieurs reprises, dans l'introduction de la *Critique de la raison pure* ou dans l'*Esthétique Transcendantale*, que « $5 + 7 = 12$ » et que « la somme des angles d'un triangle fait 180 degrés ». Ces exemples sont déjà présents dans le *Théétète* de Platon, comme je l'ai montré. Ils ont servi de soupente, ou de points de repère, à l'évolution du concept philosophique de jugement tout au long de l'histoire. Or il se trouve que, vers la fin du XIX^{ème} siècle et le début du XX^{ème} siècle, des résultats novateurs sont obtenus concernant, justement, la véracité de ces propositions.

D'une part, la révision du cinquième postulat des *Éléments* d'Euclide conduit à l'apparition des géométries non euclidiennes. Ces théories non intuitives, au sein desquelles la somme des angles d'un triangle fait tantôt plus de 180 degrés et tantôt moins, questionnent l'origine des vérités géométriques : sont-ce des conventions ? des résultats issus de l'expérience du monde ? de simples contingences de notre esprit humain ? Le mathématicien Poincaré demande à quoi aurait ressemblé notre géométrie si nous étions nés, non pas sur la terre ferme, mais dans un milieu aquatique tel que certains corps se déforment et que la lumière se réfracte ?

D'autre part, le mathématicien Hilbert donne pour projet d'axiomatiser les mathématiques, en particulier l'arithmétique, de façon à ce que des additions du type « $5 + 7 = 12$ » puissent être démontrées à partir d'un petit nombre d'axiomes. Dans un cas comme dans l'autre, l'idée s'impose que les vérités mathématiques sont des vérités logiques, analytiques, déductibles à l'aide de la raison seule. La notion d'intuition, quoique toujours présente, prend une acception différente. Il ne s'agit plus de la structure *a priori*, de l'échafaudage, sans lequel les phénomènes conscients ne sauraient exister ; mais d'une certaine aptitude à saisir l'évidence d'un axiome. Loin donc de l'idée que le monde se donne au travers des intuitions de l'esprit humain, en épousant la forme de celles-ci, l'on admet que l'intuition soit un outil pour saisir les vérités premières d'un monde qui existe en dehors de nous — et dont nous pouvons saisir la nature véritable.

En plus de faire marcher l'intuition kantienne sur la tête, ces démarches issues du champ des mathématiques ont entraîné une compréhension nouvelle, pas forcément fatale, mais quelque peu différente, de la notion de jugement. Avec l'essor du formalisme et du logicisme, le jugement redevient, comme du temps de Platon, un outil de la raison. Loin toutefois de viser à appréhender la nature des essences universelles, son rôle se cantonne à calculer, à déduire, à formuler. Cette raison calculante, faite toute de nombres, manipule les symboles, les aligne, les transforme, les efface à l'occasion, pro-

duisant des séquences nouvelles. La pensée humaine n'est plus l'expression de formes *a priori* ; c'est une astucieuse machinerie faite de rubans infinis, de têtes de lecture et de lois logiques signées « Alan Turing ». Elle prend pour la première fois la forme d'une machine à calculer. Une machine certes dantesque, à la rapidité et à l'exactitude impressionnantes, qui semble pouvoir répondre à des questions et penser de façon autonome.

Dans les sections qui suivent, je vais me concentrer plus particulièrement sur la façon dont les géométries non euclidiennes ont remis en question l'idée d'une intuition pure de l'espace. Il s'agira de déterminer si l'exemple canonique de Kant, comme quoi « la somme des angles d'un triangle fait 180 degrés », peut encore être considéré comme un jugement synthétique *a priori*. Dans la partie suivante de cet ouvrage, je me concentrerai plus spécifiquement sur la proposition arithmétique stipulant que « la somme de 7 et de 5 vaut 12 ». J'en profiterai pour montrer en quoi cela prépare le terrain à l'apparition de cet outil, aujourd'hui incontournable, qu'est l'ordinateur.

4.1 La Crise des fondements des mathématiques

Entre la fin du *XIX*^{ème} siècle et le début du *XX*^{ème} siècle eut lieu une remise en question plus ou moins fondamentale du statut ontologique et épistémologique des mathématiques. Bien qu'à l'époque elle ne fût pas toujours aperçue, elle eut une telle influence sur le développement ultérieur de la science, et en particulier de disciplines telles que la logique, la théorie des ensembles ou la théorie des catégories, qu'elle est aujourd'hui qualifiée unanimement de « crise ».

D'aucuns rapportent le début de ce phénomène aux environs de 1870^[7], au moment où le débat philosophique sur les géométries non euclidiennes atteignait son paroxysme ; d'autres préfèrent la date de 1902, en référence à la lettre que Russell fit transmettre à Frege pour lui annoncer l'existence d'un paradoxe logique dans les fondements de l'arithmétique^[8]. Mais si l'inauguration de cet événement suscite le débat, que dire de sa clôture ? Les uns pensent que la crise se termine peu après la publication des théorèmes de Gödel^[9], d'autres en font remonter bien plus loin ses conséquences. Toujours est-il que cette problématique fit couler beaucoup d'encre et mérite d'attirer notre attention.

Plusieurs controverses furent au centre de la crise. La plus connue aujourd'hui est probablement celle qui touche aux nombreux paradoxes de la théorie des ensembles — sur laquelle on cherchait tant bien que mal à faire s'asseoir l'arithmétique. Ce sont ceux de Russell, de Richard, de Burali-Forti

7. LEPAGE et FRADET, *La crise des fondements : quelle crise ?*, voir Avant-propos, pp. 1-5.

8. BAROT, *En quoi la crise des fondements des mathématiques est-elle terminée ?*, p.1.

9. *Ibidem*.

ou de Cantor^[10] Un bibliothécaire du nom de Berry synthétise la substantifique moelle de ces paradoxes dans un objet contradictoire, défini comme étant : « Le plus petit entier naturel non descriptible par une expression de quinze mots ou moins. ». L'astuce vient de ce que la phrase entre guillemets fait justement quinze mots.

Ces paradoxes logiques ont fortement marqué la postérité, passèrent presque inaperçus des mathématiciens de l'époque. Considérés comme « casse-têtes » ou de simples « jeux de langage », ils auraient réussi qu'à piquer l'ironie du célèbre mathématicien Henri Poincaré : « La logique n'est plus stérile, elle engendre des contradictions^[11]. » Ces paradoxes sont jugés résolus avec la parution, quasi simultanée, de *Set theory as based on the theory of types* (Russell) et de *Investigations in the foundations of set theory I* (Zermelo, 1908).^[12] Ce n'est que des années plus tard qu'on comprend la profondeur de leur implications en ce qui concerne le projet de bâtir les mathématiques uniquement sur des jugements analytiques.

La seconde controverse qui a marqué l'Histoire concerne le statut des mathématiques et renvoie à la querelle qui opposa le formalisme à l'intuitionnisme en les personnes de Hilbert et de Brouwer. Le premier pense que les vérités mathématiques correspondent à des objets de notre intuition, que sont les symboles, mais que cette intuition doit se déployer au travers d'un système axiomatisé de règles. Selon ce point de vue, les mathématiques formelles doivent s'accompagner d'une méta-mathématique, si possible finitiste, dont le rôle est de prouver de façon constructive la validité des règles d'inférence utilisées par les mathématiques proprement dites.

Brouwer, quant à lui, défend une vision selon laquelle aucun langage formel, aussi rigoureux soit-il, ne peut valablement exprimer l'intuition mathématique — qui est *a priori*. Il rejette la conception selon laquelle l'existence d'un objet mathématique est équivalente à sa non-contradiction, n'admettant ainsi que les preuves constructives. Pour cette même raison, il exclut des règles d'inférence le principe du tiers exclu. Ces deux positions, à bien les prendre, semblent être issues l'une comme l'autre des ruines du système kantien (dont nous verront plus loin ce qui en est responsable^[13]). Elles proposent des solutions innovantes au nouveau statut des mathématiques.

Une troisième controverse, beaucoup plus discrète celle-là, tourne autour de l'école polonaise de Lvov-Varsovie et de la pensée de Stanislaw Leśniewski (1886-1939). Ce philosophe et logicien, parallèlement aux travaux de Zermelo-Fraenkel, ouvre une troisième voie pour fonder les mathématiques. Son système est formé d'une *Ontologie*, d'une *Protothétique* et d'une *Méréologie*^[14], qui offrent respectivement une alternative au calcul des pro-

10. LEPAGE, *op. cit.*, p. 1.

11. POINCARÉ, *Les mathématiques et la logique*, p. 306.

12. LEPAGE, *loc. cit.*

13. Voir section 4.4. *Objections à Kant*.

14. La Méréologie est le premier système de Leśniewski. Il est axiomatisé, mais non

positions, au calcul des prédicats et à la théorie des ensembles. Leśniewski s'inscrit dans la lignée des logiciens, mais là où Frege se contentait d'imposer une restriction d'usage à son principe V, là où Russell introduisait la théorie des types, il souhaite créer une définition de "classe" non contradictoire et plus en accord avec l'intuition. Sa grande innovation consiste à envisager des systèmes ouverts, où les termes du langage (les noms), plutôt que d'être définis une fois pour toutes, peuvent être ajoutés indéfiniment. Le débat consistait à déterminer les liens précis entre la solution du Polonais et celles de ses contemporains.

Plutôt que de survoler ces trois controverses les unes après les autres, ou d'en approfondir une de façon privilégiée, je souhaiterais adopter une nouvelle perspective de lecture : celle-ci consiste à montrer que, tout au long de cette période mouvementée que représente la crise des fondements, s'est joué un conflit entre deux visions de la vérité mathématique qui remonte en fait aux origines de la philosophie.

D'une part, l'on trouve les irrécusables défenseurs du jugement synthétique *a priori* qui, en dépit des difficultés rencontrées par la théorie kantienne, cherchent à préserver l'apriorité sous la forme (a) d'une intuition pure ou (b) d'un platonisme mathématique. Ce sont notamment les intuitionnistes et les formalistes.

D'autre part, l'on trouve ceux qui, renonçant à l'intuition *a priori*, se tournent soit (c) du côté du jugement *a posteriori*, soit (d) du côté des jugements analytiques. L'on compte parmi eux les constructivistes, les empiristes et, enfin, les logiciens.

Pour bien faire ressortir ces lignes de force, peut-être n'est-il pas inutile de synthétiser d'abord ce que Kant et Platon ont apporté à l'épistémologie des mathématiques. Qu'est-ce qu'un objet mathématique ? Qu'est-ce qu'un raisonnement correct ? À bien considérer les choses, Kant et Platon ont ceci de commun que ce qui motive leurs systèmes est, dans une large mesure, une fascination du pouvoir des mathématiques — et la volonté d'en rendre compte^[15]. Qu'entendons-nous par là ?

Il y a dans l'exercice des mathématiques comme la révélation d'une vérité qui, sans être donnée préalablement à la raison, s'impose à elle avec la dernière nécessité. Ce qui semblait ressortir d'une évidence est tout à coup falsifié par les seules lois du raisonnement ; ce qui au contraire paraissait incertain reçoit subitement la sanction d'une démonstration. D'un côté, l'esprit a l'impression d'être libre de créer ses propres concepts, indépendamment des contraintes du monde sensible, et d'y mettre ce qu'il y souhaite. D'un autre côté, il a sans cesse le sentiment que ces concepts-là même qu'il a créés le dépassent, qu'il est impossible de les pénétrer de fond en comble, en bref,

formalisé. C'est dans l'optique de lui donner des fondements plus solides que sont venues s'adjoindre l'Ontologie et la Protothétique.

15. Le portique de l'Académie n'était-elle pas marqué de cette inscription : « Que nul n'entre ici s'il n'est géomètre ! »

qu'ils répondent de leurs propres lois.

L'objet mathématique semble parfois avoir la même autonomie qu'un objet du monde physique, à ceci près que ses contours sont dessinés avec des nombres, que sa forme est celle des essences, que sa matière est universelle. Quiconque se frotte à lui est mis en demeure de répondre à cette épineuse question : *comment rendre compte de l'apparente autonomie des objets mathématiques tout en préservant, dans le même temps, l'apparente autonomie de la raison ?*

Les *Éléments* d'Euclide ont servi de modèles aux mathématiques pendant plusieurs siècles^[16]. Cet immense édifice a révélé au monde entier qu'il est possible de découvrir les propriétés des figures géométriques en raisonnant directement à partir d'un ensemble de définitions et d'axiomes. Il est donc permis de dégager la science mathématique de ses représentations sensibles, de façon à la faire entrer tout entière dans une logique portant sur les essences. Ce faisant, les preuves acquièrent une généralité universelle. Elles deviennent vraies de tout cercle, de tout triangle, et pas uniquement de ceux qui auront été dessinés. Ce saut est inestimable. Il laisse apercevoir tout un royaume de vérité, dont il reste à l'esprit de faire la conquête.

Sans la parution des *Éléments* d'Euclide et le développement des mathématiques, peut-être que Platon n'aurait rien eu à redire au relativisme de Protagoras^[17]. Son système semble avoir été créé pour justifier 1) la supériorité du jugement essentialiste sur les apparences visibles et 2) l'autonomie des objets mathématiques, qui ne sont rien d'autre que les Idées. Ce faisant, Platon répond à la question soulevée par l'expérience mathématique. Il montre que les lois de la raison sont dotées d'une certitude que ne possède pas la connaissance sensible. Il montre en outre que l'intelligible fait face à la raison et lui est préalable. Il va même jusqu'à ramener la réalité, dans sa totalité, à ces Formes pures. Avec lui est née la position réaliste en mathématiques, nommée en son honneur le platonisme mathématique.

16. S'il est clairement le plus marquant de son temps, cet ouvrage à la date incertaine est loin d'être le premier exemple d'exposition apodictique de résultats mathématiques. Peu de temps avant lui, un savant comme Autolykos de Pitane a rédigé un traité sur les sphères en mouvement, composé de définitions et de 12 propositions. Ces résultats ne sont pas comparables aux 470 propositions des *Éléments* d'Euclide, mais laissent à penser que cette façon de rédiger les mathématiques s'insère dans une tradition plus ancienne. D'ailleurs, les *Éléments* étaient eux-mêmes la compilation d'un certain nombre de travaux antérieurs : ceux d'Hippocrate de Chios (livres I-IV et livre XI), l'arithmétique des pythagoriciens (livres VII-IX), les théories d'Eudoxe (V-VI et XII) et les résultats de Théétète sur les irrationalités quadratiques (livre X) et sur les polygones (livre XIII). Pour de plus amples détails, se référer aux deux premiers chapitres de ROSENFELD ^[17], *A History of Non-Euclidean Geometry : Evolution of the Concept of a Geometric Space*.

17. La théorie des Idées, comme nous l'avons indiqué dans la section ??, est en grande partie redevable aux pythagoriciens. Or il est connu que ces derniers plaçaient, au cœur de leur enseignement, une philosophie et une mystique des nombres, qui à son tour renvoie à une science pure de la géométrie. Ceci appuie l'hypothèse que les mathématiques furent d'un apport non négligeable à l'épistémologie de Platon.

Le renversement opéré par Kant, on l'a vu, consiste à ramener les vérités mathématiques dans le giron de l'intuition sensible, et ce au moyen des jugements synthétiques *a priori* : « Les jugements mathématiques sont tous synthétiques^[18]. » D'un même geste, il retranche à la raison la liberté de créer des concepts dans le vide, tout en maintenant le caractère universel des propositions mathématiques. Quant aux objets mathématiques eux-mêmes, ils conservent bien entendu leur autonomie et leur consistance propre. Mais l'on ne trouve plus en eux que ce que l'Entendement y a mis. Au lieu d'être les ressortissants de quelque terre étrangère, ils sont situés à même les facultés du sujet.

L'on peut donc dire qu'avec Kant, l'objet de la mathématique consiste à découvrir les lois *a priori* de ses intuitions spatio-temporelles. « L'éclatant exemple des mathématiques nous montre jusqu'où nous pouvons aller dans la connaissance *a priori* sans le secours de l'expérience. Il est vrai qu'elles ne s'occupent que d'objets et de connaissances qui peuvent être représentés dans l'intuition ; mais on néglige aisément cette circonstance, puisque l'intuition dont il s'agit ici peut être elle-même donnée *a priori*, et que, par conséquent, elle se distingue à peine d'un simple et pur concept^[19]. »

Si l'objet mathématique chez Platon est donné de façon intemporelle, il n'en va pas de même pour Kant. Aux yeux de ce dernier, l'existence mathématique, tout au moins quant à son aspect phénoménal (car la connaissance ne peut rien savoir de la chose en soi), dépend de la possibilité d'une effectuation mentale. N'existe que ce qui peut faire l'objet d'une construction dans l'intuition et dans le temps à l'aide d'une succession d'opérations finies^[20]. Les preuves d'existence non constructive, comme le fait de toute équation différentielle ordinaire qu'elle accepte une solution, ne seraient sans doute pas admises au titre d'objet mathématique aux yeux de Kant. Cette position pose les jalons d'une conception épistémologique d'importance majeure : l'intuitionnisme.

Le dénominateur commun entre ces deux points de vue, c'est la certitude de la méthode de démonstration *ad geometrico*. Pour Kant comme pour Platon, l'esprit, dès lors qu'il raisonne correctement, est capable d'atteindre le vrai de façon indubitable. Pour l'un, ce raisonnement correct tient dans le détachement du sensible ; pour l'autre, il consiste en son bon ordonnancement. Or il faut savoir que cette certitude en la méthode mathématique, déjà au XVIII^{ème} siècle, est toute relative. Mis à part le cas particulier de la géo-

18. KANT, *La Critique de la raison pure*, Introduction V, p. 58.

19. KANT, *op. cit.*, p. 52

20. Dans la théorie kantienne du schématisme, chacune des étapes intervenant dans la construction doit être finie, mais le processus lui-même peut être indéfini. Cette pensée de l'infini en puissance est, je pense, très importante chez Kant. C'est elle qui permet de comprendre les antinomies de la raison, en particulier en ce qui concerne la question de savoir si le monde est clos. Il y a en effet cette double idée que, d'une part, les limites du monde reculent sans cesse devant la raison et que, d'autre part, il faut nécessairement leur trouver un terme.

métrie euclidienne, les disciplines mathématiques souffrent d'un manque de rigueur et de fondements^[21]. Il n'existe pas d'axiomatisation pour l'arithmétique ou l'analyse. De plus, les règles de raisonnement ne sont pas toujours clairement définies. L'on mêle des déductions logiques finitistes aux éléments infinitésimaux. L'infini en puissance se cache sous les passages à la limite. Mais en dépit de ces imperfections, l'on prête à tous les résultats mathématiques le même prestige qu'aux *Éléments* d'Euclide. On le fait, du moins, jusqu'au jour où la géométrie, elle-même, finit par être ébranlée. C'est de ce séisme intellectuel dont il va être question dans les pages suivantes.

4.2 Le Livre de Dieu et les Mondes parallèles

Il est certain livres dont la pureté, l'élégance et la profondeur ont eu un tel impact sur l'histoire de l'humanité que, si on ignorait leur auteur, on aurait aisément vu un livre de Dieu. Aux rangs de ces ouvrages exceptionnels figurent, sans aucun doute, les *Éléments* d'Euclide. Le nombre de rééditions de ce traité de mathématiques, composé initialement de treize livres distincts, n'est d'ailleurs surpassé que par la Bible.

Or ce qui est ébranlé par les progrès mathématiques du *XIX^{ème}* siècle, c'est précisément la toute-puissance de cette Bible mathématique. La désillusion, notons-le bien, ne s'est pas faite en un jour. Depuis les Grecs déjà, l'un des postulats d'Euclide, connu sous le nom du postulat de la parallèle, avait attiré l'attention pour son manque d'évidence. Il paraissait moins sûr, moins intuitif. Nombre de mathématiciens anciens, comme Proclus, tentèrent de le prouver sur la base des quatre premiers axiomes, mais sans succès. Schopenhauer proposa d'y voir un jugement synthétique *a priori*. Or bientôt, il s'avéra que ce postulat n'était ni démontrable, ni même nécessaire, et que d'autres géométries, toutes plus étonnantes les unes que les autres, pouvaient naître de son abandon. L'on se rendit compte que le raisonnement pouvait mener à des vérités géométriques qui transcendent l'intuition.

Le postulat des parallèles figure dans le livre I d'Euclide. On le trouve exprimé après un ensemble de 35 définitions concernant les points, les lignes, les surfaces. Il accompagne quatre autres postulats, de nature bien plus "intuitive". Je les donne ici tels qu'ils ont été énoncés par Euclide^[22], mais il est clair que, depuis lors, ils ont été exprimés de façon beaucoup plus formelle^[23]:

1. L'on est capable de tracer une unique droite à partir de deux points quelconques.

21. Voir à cet effet l'excellent article de vulgarisation de MORRIS KLINE intitulé *Les Fondements des mathématiques*.

22. En suivant la traduction de Peyrard depuis PERRIN, *Les géométries non euclidiennes*, p.1.

23. En fait, ces 5 postulats ne sont pas complets pour la géométrie euclidienne. Aujourd'hui, ils sont au nombre de douze

2. L'on peut prolonger continûment un segment pour en faire une ligne droite.
3. Il est possible de tracer un cercle de n'importe quel rayon en prenant n'importe quel point pour centre.
4. Tous les angles droits sont égaux les uns aux autres.
5. Pour tout point A du plan et pour toute droite D, il existe une et une seule droite parallèle à D passant par A^[24].

Selon l'historien russe Boris Rosenfeld, Euclide aurait introduit le cinquième postulat de façon à prévenir une série d'erreurs commises en son temps^[25]. Ce serait donc une véritable innovation de sa part. Plus encore, cela signifierait qu'il s'était peut-être rendu compte que ce postulat, à l'allure pourtant simple, n'était pas une conséquence logique des quatre précédents. Tout se passe d'ailleurs comme si Euclide cherchait à prouver un maximum de théorèmes sans en faire usage. C'est seulement à partir de la proposition 29 qu'il se trouve obligé d'y recourir. Cette proposition stipule que (voir figure ^[??]) :

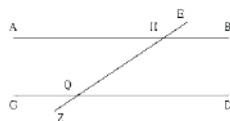


FIGURE 4.2 – Proposition XXIX

Proposition 2 (Proposition XXIX (Livre II)). *Une droite qui tombe sur deux droites parallèles fait les angles alternes égaux entre eux, l'angle extérieur égal à l'angle intérieur opposé et placé du même côté, et les angles intérieurs placés du même côté égaux à deux droits.*

À partir de là, le postulat fait des apparitions de plus en plus fréquentes, et notamment dans les démonstrations des propriétés suivantes :

24. Cette formulation intuitive du cinquième postulat n'est pas celle que l'on trouve dans le livre d'Euclide, mais sous la plume de Proclus. En fait, Proclus ne l'énonce même pas comme tel, mais le suppose dans une démonstration. L'énoncé original d'Euclide est le suivant : « Si une droite tombant sur deux droites fait les angles intérieurs du même côté plus petits que deux droits, ces droites, prolongées à l'infini, se rencontreront du côté où les angles sont plus petits que deux droits. »

25. ROSENFELD ^[171], *op. cit.*, p. 38. Son argument se base sur une lecture d'Aristote : « En parlant de l'erreur appelée *pétition de principe*, Aristote fait remarquer que : "C'est ce qui se produit avec ceux qui pensent qu'ils décrivent des lignes parallèles, alors qu'ils supposent inconsciemment des choses qu'il n'est pas possible de démontrer en l'absence des lignes parallèles" » (cité des *Premiers Analytiques* par ROSENFELD, *op. cit.*, p. 38).

- La relation de parallélisme est transitive : si une droite D_1 est parallèle à une droite D_2 et que cette droite D_2 est parallèle à une droite D_3 , alors D_1 est parallèle à D_3 . ;
- Si les droites D et D' sont perpendiculaires, D est également perpendiculaire à toutes les parallèles de D' ;
- La somme des angles d'un triangle est égale à 180° ;

L'idée de fonder ces propriétés sur un postulat qui, à son tour, s'apparente plus à un théorème qu'à une "évidence incontestable" préoccupe beaucoup les géomètres. Jusqu'au $XIX^{\text{ème}}$ siècle, de nombreux mathématiciens anciens^[26], arabes^[27] et de l'Europe renaissantiste^[28] tentent en vain de fournir une preuve du postulat^[29]. La plupart des tentatives, lorsqu'elles ne dissimulent pas une pétition de principe, se ramènent à deux cas : soit une nouvelle définition de la parallèle était donnée, soit un postulat équivalent à celui d'Euclide, mais plus "intuitif", est fourni en guise de remplacement. Un exemple est la définition de Posidonios, selon laquelle une parallèle D' à une droite D est *le lieu des points équidistants de D et située du même côté de cette droite*.

Le postulat d'Euclide en serait resté là, s'il ne s'était pas présenté un homme du nom de Saccheri. Ce mathématicien italien, décrit comme un enfant précoce, et ordonné prêtre en 1694, s'entête à donner une preuve définitive au postulat. Au lieu de chercher des preuves directes, il contourne adroitement le problème. Il suppose que le postulat est faux et tente de montrer que cela entre en contradiction avec les autres postulats (le deuxième en particulier). Cette stratégie est connue sous le nom de démonstration par l'absurde. Elle conduit notamment à la construction de « carrés » avec des côtés courbes, ou à des triangles dont la somme des angles n'est pas égale à 180° ^[30]. Saccheri n'arrive pas à obtenir les contradictions voulues. Tout se passe comme si la logique lui refusait toute coopération. Il mourra deux mois après la publication de ses travaux, dans l'ignorance quasi totale.

Saccheri n'était pas le premier à avoir procédé à des raisonnements par l'absurde. On trouve des idées similaires dans les travaux arabes de Omar Khayyam^[31]. En fait, la suspension du postulat d'Euclide, bien que non intuitive, n'a rien d'absurde. Elle peut conduire à deux types bien différents de géométries logiquement cohérentes : le premier est celui des géométries dites elliptiques, où la somme des angles d'un triangle est strictement supé-

26. Archimède, Posidonios, Ptolémée, Proclus...

27. Al Jawhari, Thabit Ibn Qurra, Ibn al-Haytam, Khayyam, Husam al-Din al-Salar, Nasir al-Din al-Tusi, al-Hanafi, al-Abhari, al-Maghribi...

28. Gerson et Alfonso, Clavius, Cataldi, Giodano, Wallis, Saccheri, Clairaut, Lambert, Bertrand, Legendre, Farkas Bolyai...

29. Pour un historique complet de ces tentatives, se référer à ROSENFELD, pp. 41-109.

30. ROSENFELD [171], *op. cit.*, pp. 98-99.

31. KHAYYAM OMAR, *Commentary on the Difficulties Concerning the Postulates of Euclid's Elements*

rieure à 180° et où il n'existe pas de droites parallèles ; le second est celui des géométries hyperboliques, où la somme des angles d'un triangle est strictement inférieure à 180° et où il est possible de tracer, pour toute droite D quelconque, une infinité de parallèles à D passant par un même point (voir figure ??). Ce n'est que dans le milieu du $XIX^{\text{ème}}$ siècle que les mathématiciens comme Beltrami, Lambert, Lobatchevski, Bolyai, Riemann, Gauss et Klein³² vont accepter la possibilité de ces théories et en formaliser les notions.

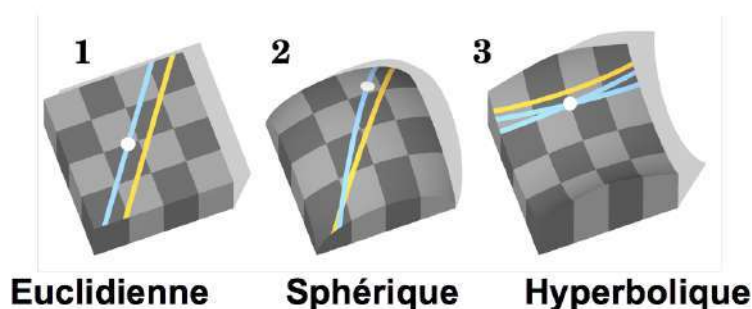


FIGURE 4.3 – Soit une droite D (en vert) et un point A (en blanc). En géométrie euclidienne, il existe une unique droite parallèle à D passant par A. En géométrie elliptique, il n'en existe aucune ; en géométrie hyperbolique, une infinité.

Toutes les géométries non euclidiennes dépendent de la façon dont l'on définit la longueur d'une courbe, car de ces définitions, il pourra venir que le chemin le plus court n'est pas donné par la ligne "droite" au sens euclidien. Mais dans la mesure où le rejet du postulat des parallèles, loin de conduire à des absurdités, permet au contraire de définir des théories aussi cohérentes qu'originales, il paraît légitime de se demander combien de géométries non euclidiennes sont possibles. Pour nous aiguiller sur cette question, l'on peut recourir à une application du théorème de Lie, stipulant que le nombre de géométries possibles est toujours borné pour peu que les conditions suivantes soient satisfaites³³ :

- L'espace est à n dimensions.
- Le mouvement d'une figure invariable est possible.
- Il faut p conditions pour déterminer la position de cette figure dans l'espace.

32. *Ibidem*, pp. 289 - 297 pour les géométries elliptiques et pp. 206 - 221 pour la géométrie hyperbolique.

33. Cette application vulgarisée du théorème de Lie est donnée par POINCARÉ dans *La Science et l'Hypothèse*, p. 63.

L'hypothèse 2 stipule qu'un solide en déplacement dans l'espace ne doit pas subir de déformation au cours de son déplacement, et ce quelle que soit sa trajectoire. En levant cette hypothèse, il est aisé de construire une infinité de géométries, comme Riemann en a donné la preuve³⁴. L'on distingue aujourd'hui les géométries algébriques, différentielles et analytiques. Pour retracer l'impact que ces géométries ont eu sur l'épistémologie des mathématiques et sur le statut du jugement synthétique *a priori*, je vais me limiter à une brève description des deux grandes classes de géométries qui ont marqué la moitié du *XIX^{ème}* siècle : les géométries elliptiques et hyperboliques.

4.2.1 Géométries sphériques et elliptiques

Considérons une sphère et faisons passer un plan par son centre. L'intersection définit ce que l'on appelle un *grand cercle*, qui peut être identifié à la "droite" de la géométrie euclidienne. Les "segments" sont les arcs de ces grands cercles. On peut les utiliser pour tracer des triangles, des carrés, etc. La géométrie ainsi obtenue est la géométrie dite sphérique.

Il est aisé de voir qu'entre deux points A et B non antipodaux, il est possible de faire passer un seul grand cercle. Mais dans le cas particulier où A et B sont des antipodes, il devient possible d'en faire passer une infinité. Ce résultat viole le premier axiome. La géométrie sphérique n'est donc pas euclidienne.

Le troisième axiome, lui aussi, rencontre des limites d'application : s'il reste possible de tracer des cercles depuis n'importe quel centre, il est en revanche inadmissible de leur donner un rayon supérieur à celui de la sphère. Mais la question intéressante pour notre propos consiste à se demander ce que devient le cinquième postulat ? Sous la forme que lui a donnée Euclide, il est tout à fait valide. Et pour cause : toutes les droites se coupent sur une sphère. Mais si l'on considère la forme de Proclus, il en va bien différemment. Cette version, rappelons-le, dit qu'il n'existe qu'une seule droite passant par un point et parallèle à une droite donnée. Comme il n'existe en fait aucune parallèle en géométrie sphérique, l'axiome doit être changé de la manière suivante :

Postulat 1 (Proposition XXIX (Livre II)). *Pour tout point A du plan et pour toute droite D , il n'existe pas de droite parallèle à D passant par A .*

Les géométries sphériques (*sphericae*) ont fait l'objet de nombreuses réflexions et traités tout au long de l'Antiquité. La principale motivation de ces études était l'observation des corps célestes et la mise au point d'une astronomie. Dans son *Almageste*, Ptolémée consigne un grand nombre d'observations ayant été faites par les Babyloniens au *VIII^{ème}* siècle avant notre ère. C'est à eux que l'on doit la division de l'écliptique en douze signes du

34. *ibidem*.

zodiaque, de 30° chacun, ainsi que la division sexagésimale des degrés en minutes et en secondes^[35]. Il faut attendre le $IV^{\text{ème}}$ siècle avant J.-C. pour que les Grecs prennent connaissance de cette tradition astronomique et commencent à la développer.

Des propriétés plus générales concernant les triangles sphériques sont prouvées au $XVI^{\text{ème}}$ siècle. Dans un article publié en 1629, le mathématicien hollandais Albert Girard (1595 - 1632) démontre qu'au plus grande est la surface d'un triangle sphérique, au plus grande est la somme de ses angles^[36]. Plus précisément, il montre que la différence entre la somme des angles du triangle et 180° (quantité que l'on appelle *l'excès angulaire*) est proportionnelle à l'aire du triangle. Il parvient même à étendre ce résultat à n'importe quel type de polygone sphérique. Il avait conscience que ses preuves n'étaient pas tout à fait rigoureuses, car elles nécessitaient l'introduction du calcul infinitésimal. Euler en donnera une démonstration complète, simple et élégante^[37]. Il donnera aussi des expressions trigonométriques reliant l'aire d'un triangle à la longueur de ses côtés, contribuant grandement au développement de la trigonométrie sphérique de Ptolémée et de Copernic.

Toutes ces géométries sont parfaitement compatibles avec notre intuition. Elles traitent en effet des propriétés de polygones situés à la surface d'une sphère, qu'il est parfaitement possible de se représenter dans un espace euclidien. Il pourrait être intéressant de se demander s'il existe des géométries sphériques à trois dimensions ? Ou s'il en existe qui satisfont les 4 premiers postulats d'Euclide, mais pas celui des parallèles ?^[38] Il est difficile de répondre à ces questions à l'aide de notre intuition spatiale, mais il est possible de s'en faire une idée par une démarche hypothético-déductive.

Le mathématicien Riemann le premier entreprend cette démarche, aux alentours de 1956^[39]. Le résultat est l'obtention d'un vaste ensemble de propositions mathématiques logiquement justifiées, que l'on nomme la géométrie elliptique. Cette théorie admet un certain nombre de modèles, ou encore, de représentations spatiales. Dans chacun de ces modèles, les objets de la théorie sont mis en relation avec des objets géométriques intuitifs (droites, courbes, sphères...) et dotés de règles décrivant leurs comportements.

Le modèle le plus simple consiste à se représenter une lentille sphérique telle que chaque paire d'antipodes représente un "point" de notre géométrie. Pour le dire autrement, on ne garde qu'un seul des hémisphères de la sphère : le second est sa copie dans un miroir. Les "droites" de ce modèle

35. ALMAGESTE [?], p. 166

36. ROSENFELD [171], *op. cit.*, p. 27-28.

37. *Ibidem*, p. 32.

38. Notons que dans la formulation axiomatique contemporaine de la géométrie euclidienne, l'on dénombre non pas 5 postulats, mais 12. Le $XII^{\text{ème}}$ correspond au postulat des parallèles. C'est le seul à ne pas être respecté par la géométrie elliptique.

39. ROSENFELD [171], *op. cit.*, p. 289

sont les grands cercles de la sphère (les cercles de diamètre maximal). Si l'on considère un point M quelconque de la sphère et une "droite" D , l'on voit qu'il n'existe aucune parallèle à D passant par M . De plus, le résultat selon lequel la somme des angles d'un triangle est supérieure à 180° est conservé. L'on appelle ce modèle la géométrie elliptique sphérique, pour éviter de le confondre avec la géométrie sphérique en deux dimensions^[40].

4.2.2 Géométries hyperboliques

Peu avant l'introduction des géométries elliptiques, une alternative intéressante au postulat des parallèles avait déjà été pressentie par Lobatchevski, Gauss et Bolyai. Celle-ci consiste à partir du constat suivant : dans les géométries elliptiques, la somme des angles d'un triangle est supérieure à 180° . Dans les géométries euclidiennes, elle est égale à 180° . N'existerait-il pas une géométrie pour laquelle cette somme est inférieure ? Dans un article intitulé *Sur les principes de la géométrie*, Lobatchevski exprime cette idée :

« Il reste à faire l'hypothèse que cette somme est soit $= \pi$, soit $< \pi$. Les deux sont envisageables sans contradiction, et cela donne naissance à deux Géométries : l'une, jusqu'à présent préférée en raison de sa simplicité, s'accorde en fait à toutes les mesures ; l'autre, qui est imaginaire et qui, en raison de sa plus grande généralité, rend les calculs plus compliqués, admet une dépendance entre la taille des côtés du triangle et la somme de ses angles^[41]. »

L'une des propriétés de cette géométrie imaginaire, c'est que, de façon tout à fait analogue à ce que l'on observait dans l'espace elliptique, la somme des angles d'un triangle s'écarte le plus de 180° lorsque ses côtés s'accroissent. Ici évidemment, l'écart est négatif. Cela dit, bien d'autres propriétés moins "intuitives" peuvent être observées. Parmi celles-ci, relevons les plus importantes^[42] :

- **Cercles limites** : Un cercle dont le rayon tend vers l'infini ne se comporte pas comme une droite, mais comme une courbe de genre 2 (c.-à-d. une courbe en deux dimensions), spécifique à la géométrie de Lobatchevski, et que l'on nomme un *cercle limite*.

40. Notons que, en géométrie sphérique, il est possible de faire passer une infinité de droites par les deux points antipodaux d'une sphère, ce qui contredit le premier postulat d'Euclide. En géométrie elliptique sphérique, par contre, les points antipodaux sont confondus, ce qui permet de contourner cette situation et de conserver le premier postulat.

41. ROSENFELD [171], *op. cit.*, p. 207

42. Les 4 propriétés sont extraites de l'article *Sur les principes de la géométrie* de Lobatchevski, selon les citations qui en sont faites dans ROSENFELD [171], p. 207.

- **Sphères limites** : Pareillement, une sphère dont le rayon tend vers l'infini se rapprochera d'un objet de genre 3 que l'on appelle une *sphère limite*. Lorsqu'on coupe cette sphère avec un plan, l'on trouve soit un cercle, soit un cercle limite.
- **"Droites"** : La géométrie dans une sphère limite est "analogue" à la géométrie euclidienne dans le plan. Tous les cercles limites de la sphère limite peuvent être interprétés comme les "droites" de l'espace euclidien⁴³. De plus, les angles séparant ces "droites" peuvent être mesurés en considérant les angles des plans dans lesquels ces cercles limites sont compris.
- **Parallèles multiples** : Le cinquième postulat d'Euclide dans une sphère limite est violé. Pour toute "droite" D et pour tout point P extérieur à D, il existe une infinité de "droites" passant par P et parallèles à D.

Si ces propriétés peuvent sembler difficiles à se représenter, c'est précisément qu'elles ne s'accommodent pas avec notre intuition spatiale. Il est cependant possible d'en donner des interprétations visuelles, soit en projetant cet espace dans un cercle, soit par projection dans un demi-plan infini. Dans son ouvrage de vulgarisation intitulé *La Science et l'Hypothèse*, Poincaré propose l'analogie suivante :

« Supposons un monde renfermé dans une grande sphère et soumis aux lois suivantes : la température n'y est pas uniforme ; elle est maximum au centre et elle diminue à mesure qu'on s'en éloigne, pour se réduire au zéro absolu quand on atteint la sphère où ce monde est renfermé. Je précise davantage la loi suivant laquelle varie cette température. Soit R le rayon de la sphère limite ; soit r la distance du point considéré au centre de cette sphère. La température absolue sera proportionnelle à $R^2 - r^2$. Un objet mobile deviendra alors de plus en plus petit, à mesure qu'on se rapprochera du cercle limite. Observons d'abord que si ce monde est limité au point de vue de notre géométrie habituelle, il paraîtra infini à ses habitants. Quand ceux-ci en effet veulent se rapprocher du cercle limite, ils se refroidissent et deviennent de plus en plus petits. Leurs pas deviennent donc de plus en plus petits, de sorte qu'ils ne peuvent jamais atteindre ce cercle limite⁴⁴. »

Les habitants de ce monde ne sont pas capables de réaliser que leur taille se modifie en parcourant l'espace, puisque leurs instruments de mesure

43. Ils respectent tous les postulats d'Euclide, à l'exception du cinquième.

44. POINCARÉ, *La Science et l'Hypothèse*, p. 84-85.

suivent la même variation. Il est possible de montrer que, pour eux, le plus court chemin menant d'un point à un autre est une courbe. Un individu qui partirait d'un point A situé tout près de la surface de son monde sphérique et qui voudrait se rendre à un autre point B situé, lui aussi, tout près de la surface devrait effectuer un arc de cercle. Ces arcs de cercle correspondent aux droites de la géométrie classique. On les appelle *cercles limites*.

Si nous projetons cet espace sur un disque, nous obtenons la représentation appelée cercle de Poincaré. Sur la figure (cf. [??](#)) sont indiqués trois droites de couleur bleue, rouge et verte, et représentent le plus court chemin entre les deux points qu'ils relient chacun. La dénomination de cercle *limite* prend maintenant tout son sens, puisque ces cercles coupent la surface du disque. Or, selon la description que nous avons donnée de ce monde, il est normalement impossible d'en atteindre la limite. De plus, nous voyons que par un même point peuvent passer deux droites (le rouge et le bleu) qui ne croisent pas une troisième (le vert). On peut dire de ces droites rouge et bleu qu'elles sont deux parallèles distinctes de la droite verte — et passant par un même point. La notion de parallélisme, en effet, est synonyme de « non intersection » Nous obtenons ainsi une contradiction du cinquième postulat d'Euclide, qui disait que par un point donné ne pouvait passer qu'une seule droite parallèle à une droite donnée. Une analyse plus rigoureuse montre que cette interprétation est formellement valide.

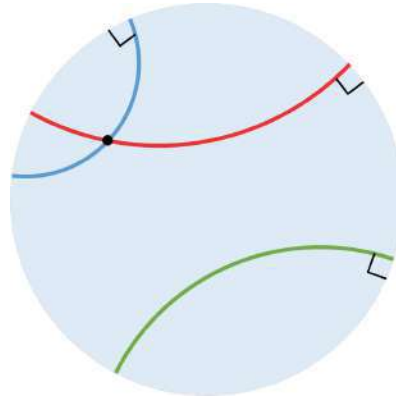


FIGURE 4.4 – Disque de Poincaré

4.3 Une Géométrie imaginaire

Au début du XIX^{ème} siècle, on savait que le monde n'était pas plat, mais nul n'aurait eu l'audace de penser que l'Espace puisse se courber sous les étoiles. Mettez-vous un instant dans la tête d'un astronome face à sa lunette d'observation. N'est-il pas vrai qu'une droite, tirée entre deux points lumineux du cosmos, troue le vide de façon rectiligne ? L'idée qu'une telle

droite dévie en suivant les replis invisibles du ciel ne frôlait pas l'imagination des rationalistes en robe de chambre. Le miracle, pourtant, s'est produit. La géométrie euclidienne, abandonnant son cinquième postulat comme on jette un sac de sable par la nacelle d'une montgolfière, a engendré des monstres. Et tandis que notre savoir s'élève toujours plus haut vers les cieux sempiternels, une question flotte sur toutes les lèvres : quel rapport existe-t-il entre les théories géométriques non euclidiennes et cet espace euclidien intuitif dont nous faisons chaque jour l'expérience ?

Les termes choisis par Lobatchevski pour désigner sa géométrie n'ont pas été laissés au hasard. En l'appelant *géométrie imaginaire*, l'auteur veut suggérer que le rapport que sa géométrie entretient avec la géométrie euclidienne est analogue au rapport existant entre les nombres imaginaires et les nombres réels. Ce sont deux axes perpendiculaires, deux dimensions complémentaires. Tant qu'on maintient le paramètre de courbure à zéro, la géométrie non euclidienne se comporte à s'y méprendre comme la géométrie des *Éléments* d'Euclide. La somme des angles d'un triangle fait 180° et une seule droite parallèle à une deuxième passe par un point donné. Or lorsqu'on lui donne des valeurs positives ou négatives, à ce paramètre, les choses changent ! L'espace devient tantôt elliptique, tantôt hyperbolique ; les triangles font soit moins de 180° , soit plus, mais ils refusent de s'en tenir à ce qu'on croyait connaître d'eux. La structure spatiale ainsi conçue doit être vue comme une généralisation de la géométrie usuelle. Il y a plus : on se rend compte assez rapidement que la géométrie imaginaire — tout comme les nombres imaginaires — peut être utilisée pour résoudre des classes d'intégrales réelles jusqu'alors insolubles.⁴⁵

La vraie question qui doit être posée, dès lors, est celle de savoir si cette géométrie aux propriétés étonnantes ne serait pas, en un certain sens, plus réelle que notre espace euclidien ? Se pourrait-il que notre espace ait cette structure étonnante, sans pourtant que nous nous en apercevions ? Lobatchevski s'est très tôt intéressé à cette question. Il montre que pour des figures géométriques très petites, telles que celles avec lesquelles nous sommes familiers, les propriétés de l'espace imaginaire ne se font pas ressentir. Dans une telle situation, la somme des angles d'un triangle vaut pratiquement 180° et la lumière semble décrire des trajectoires euclidiennes. Les différences ne se manifesteraient que si l'on pouvait envisager des trajectoires ou des triangles de très grandes tailles.

Désireux de vérifier son hypothèse, Lobatchevski entreprend un certain nombre de mesures astronomiques. Se référant aux données en vigueur à son époque, il évalue les angles d'un triangle dont le premier sommet coïncide avec l'étoile de Sirius, et dont les deux autres sont des antipodes de la Terre. Un écart de 0,00032 seconde d'arc est obtenu, ce qui le conduit à la conclusion sceptique qu'il est impossible de découvrir expérimentalement

45. ROSENFELD, *op. cit.*, p. 210.

la géométrie de notre univers^[46]. La seule présence d'un écart, aussi petit soit-il, ouvre la possibilité à un espace non euclidien, si bien qu'il faudrait jouir d'une mesure "parfaite" pour écarter l'hypothèse à coup sûr.

Les appareils de mesure contemporains montrent que l'écart est encore cent fois inférieur à celui obtenu par Lobatchevski^[47]. L'espace qu'il a découvert, et qui porte aujourd'hui son nom, a cependant reçu un bel avenir dans le progrès scientifique. Loin d'être regardé comme une curiosité, ou d'être écarté de façon dogmatique, il est intégré à titre essentiel dans un grand nombre des cosmologies du *XX^{ème}* siècle. La théorie de Friedman-Lemaître, par exemple, prédit des cycles successifs de dilatation et de condensation de l'Univers. Le rayon de courbure pourrait alternativement devenir positif, auquel cas l'Univers adopte une géométrie sphérique, puis négatif, signifiant qu'il devient hyperbolique ; en passant subrepticement par zéro — qui correspond au cas particulier de la géométrie euclidienne.

4.3.1 Le Constructivisme de Lobatchevski

Cette remise en cause ontologique s'accompagne d'une révision de la notion de l'espace. Jusqu'à présent, les vérités mathématiques étaient, dans le prolongement kantien, pensées comme des jugements synthétiques *a priori*. Or Lobatchevski affirme que « Les concepts doivent être appris par les sens — ceux qui sont innés ne doivent pas être crus^[48]. »

De son point de vue, les idées géométriques ne sont pas données de façon apriorique grâce à l'exercice d'une quelconque intuition. Ce sont bien au contraire des constructions de l'esprit humain issues de l'expérience. Dans sa *Géométrie* (1823), il explique par quel procédé sont abstraites les notions géométriques^[49]. Tout d'abord, dit-il, l'esprit ne garde de l'expérience qu'il fait des corps solides que leur propriété d'avoir une extension. C'est sur ce royaume qu'il bâtit l'idée de l'espace, et ce de façon itérative. Qu'est-ce qu'il faut entendre par là ?

L'esprit se représente aisément que deux corps étendus A et B peuvent entrer en contact l'un de l'autre pour venir former un troisième corps C. Ce dernier est conçu comme un tout distinct de A et B, bien que les incluant. En répétant cette expérience avec tous les corps de la nature, on peut se représenter le monde entier comme un seul et même corps. De cette abstraction naît l'idée de l'espace. À partir de là, l'esprit organise l'espace, en définissant des sections principales (les axes), la notion de transformations invariables (translations, rotations...), de positions relatives et de distance. Cela lui permet alors de définir la sphère comme le lieu des points équidistants d'un même centre, le plan comme le lieu des points équidistants de

46. *Ibidem*, p. 208.

47. *Ibidem*.

48. Cité depuis *ibidem*, 209.

49. *Ibidem*, pp. 210-212.

deux points, etc.

« Strictement parlant, tout ce que nous savons de la nature est le mouvement, sans lequel l'impression des sens est impossible. Par suite, toutes les autres notions, comme celles de la Géométrie, sont artificiellement conçues par notre esprit, en ce qu'elles sont abstraites des propriétés du mouvement : l'espace lui-même, séparément, n'existe pas pour nous. Cela étant, aucune contradiction ne peut survenir dans l'esprit si nous admettons que certaines forces de la nature suivent une géométrie particulière, et que d'autres suivent leurs propres géométries⁵⁰. »

4.3.2 Le Consensualisme de Poincaré

Lobatchevski n'est pas le seul à remettre en doute l'épistémologie kantienne. Des mathématiciens aussi éminents que Poincaré sont eux aussi gagnés de scepticisme :

Les axiomes géométriques ne sont donc ni des jugements synthétiques *a priori* ni des faits expérimentaux. Ce sont des conventions ; notre choix, parmi toutes les conventions possibles, est guidé par des faits expérimentaux ; mais il reste libre et n'est limité que par la nécessité d'éviter toute contradiction. C'est ainsi que les postulats peuvent rester rigoureusement vrais quand même les lois expérimentales qui ont déterminé leur adoption ne sont qu'approximatives⁵¹.

Cette position exprime une sorte de conventionnalisme modéré, ou de nominalisme. Aux yeux de Poincaré, aucune géométrie ne peut ontologiquement ou épistémologiquement prévaloir sur les autres. L'existence en mathématiques n'est rien d'autre que la non-contradiction, si bien que l'on peut légitimement dire de toutes ces géométries qu'elles existent. Quant à savoir quelle géométrie décrit le mieux notre monde physique ? En l'état de nos connaissances, cette question-là est le plus souvent indécidable.

Cela veut-il dire que toutes les géométries se valent ? Bien au contraire. Sur le plan pratique, Poincaré avance deux d'arguments en faveur de la géométrie euclidienne⁵² :

- **Simplicité** : D'une part, il s'agit de la représentation la plus simple, techniquement parlant. Les formules de trigonométrie sphérique y sont incomparablement plus courtes et élégantes. Certes, il peut exister certaines intégrales pour lesquelles la représentation de Lobatchevski est plus commode, mais pour la grande majorité des calculs

50. *Ibidem*, p. 210. La traduction est de moi.

51. POINCARÉ ^[155], *op. cit.*, p. 67.

52. *Ibidem*, p.30.

et des modèles que le mathématicien est amené à utiliser, c'est la plus efficace. Poincaré précise : « et elle n'est pas telle seulement par suite de nos habitudes d'esprit ou de je ne sais quelle intuition directe que nous aurions de l'espace euclidien ; elle est la plus simple en soi, de même qu'un polynôme du premier degré est plus simple qu'un polynôme du second degré⁵³. »

- **Accord avec l'expérience** : À notre échelle de matière, la géométrie euclidienne s'accommode bien avec les objets physiques qui nous entourent, et permet de rendre compte de leurs propriétés, de leurs déplacements, etc. Poincaré conçoit qu'il puisse en aller tout autrement si les corps n'étaient pas pour la majorité solides, mais seulement liquides ou gazeux, ou s'ils se déformaient de façon significative lors de leurs déplacements. Dans ce cas, il aurait pu se faire que d'autres géométries aient été plus appropriées.

4.4 Kant se retourne dans la tombe

Comme le suggèrent les opinions de Poincaré et de Lobatchevski, les nouvelles géométries portent un coup rude à l'idée qu'il existe des structures *a priori* de la pensée mathématique. En quel sens exactement les arguments de Kant sont-ils mis en danger ? Que deviennent les jugements synthétiques *a priori* ? Pour répondre à ces questions, il peut être utile de rappeler tout d'abord la position de Kant. À ses yeux, comme il a été dit dans le chapitre 2.3, l'espace ne représente ni une pure idéalité, ni une pure réalité. Il faut récuser aussi bien la position qui consiste à y voir un en-soi doté de propriétés intrinsèques que la position consistant à y voir une construction de l'esprit. Je vais analyser les arguments qu'il donne en défaveur de ces deux thèses et les discuter dans le cadre des géométries non euclidiennes.

4.4.1 L'Espace est-il une chose en soi ?

Le premier argument de Kant sur cette question, qui est sans doute le plus important, consiste à dire que l'espace est ce grâce à quoi les choses peuvent être pensées hors de soi. En effet, c'est par lui que peut être conçu le fait que deux choses, identiques ou non quant à leurs propriétés internes, se tiennent l'une et l'autre en des lieux différents. L'espace ne saurait donc être le produit d'une opération empirique. Il est la condition même sous laquelle les objets peuvent être représentés. Cet argument en tant que tel est indépendant de la géométrie utilisée et n'est donc pas remis en question par le développement des géométries non euclidiennes. D'aucuns diraient peut-être que le fait d'avoir une intuition de l'espace n'empêche pas, par

53. *Ibidem*.

ailleurs, l'espace d'être aussi une propriété du monde ? Je laisse là ce débat, pour me concentrer sur un deuxième argument.

Dans les *Prolégomènes à toute métaphysique future*, Kant attire notre attention sur le statut épistémologique des énantiomères^[54]. Ces objets sont les images l'un de l'autre dans un miroir et occupent donc une position analogue à la position que ma main droite occupe par rapport à ma main gauche. Ils ne diffèrent par aucune de leurs propriétés internes, si bien que la description complète de l'un nous renseigne parfaitement sur le second. En dépit de cela, l'on ne peut pas dire qu'un gant droit est identique à un gant gauche. Cet exemple exhibe selon Kant une différence qui n'est ni purement *interne* aux objets, ni non plus purement *externe* (ou relationnelle). Des propriétés comme "être à gauche de ceci" ou "être au-dessus de cela" sont en effet simplement relationnelles. Il suffit de changer la disposition des objets pour inverser ces rapports : ce qui est à droite peut être mis à gauche, ce qui est au-dessous peut être mis au-dessus, etc. Or dans le cas des gants, aucune symétrie ne permet de mettre un gant droit à la main gauche. C'est une propriété en quelque sorte "intrinsèque", mais qui n'apparaît que lorsque l'objet est considéré en référence au tout de l'espace. Cette propriété est appelée la chiralité.

Qu'est-ce que cela signifie aux yeux de Kant ? Que les objets de notre expérience sensible ne sont pas des en soi, sans quoi l'on pourrait les décrire uniquement en termes de propriétés internes ou externes. La chose en soi existe en effet indépendamment de la façon dont elle nous apparaît. Or ici, il semblerait que nous n'ayons devant nous que des manifestations des choses en soi, préalablement structurées par la forme *a priori* de notre sensibilité. C'est parce que les objets se donnent dans une intuition *a priori* qu'apparaît la chiralité. Cette propriété est le signe que les objets ne peuvent être connus que dans un espace, et que cet espace n'est lui-même que le mode de réception des objets sensibles :

Nous ne pouvons donc rendre intelligible par aucune notion particulière la différence des choses semblables et égales, mais cependant asymétriques (*incongruentes*) ; nous ne le pouvons que par le rapport à la main droite et à la gauche, qui regarde immédiatement une intuition^[55].

L'irruption des géométries non euclidiennes met cet argument en difficulté. Dans certains espaces, en effet, il est possible de créer des symétries inversant la chiralité. Un philosophe allemand du nom de Wittgenstein écrit à juste titre que « le gant de la main droite pourrait être mis sur une main gauche si l'on pouvait le retourner dans un espace de quatrième dimension^[56]. » En fait, un phénomène semblable apparaît déjà en première

54. KANT, *Prolégomènes à toute métaphysique future*, § XIII.

55. *Ibidem*.

56. WITTGENSTEIN, *Tractatus* [207], 6.36111.

dimension. Les deux segments a et b représentés sur la figure 4.5 sont des énantiomères dans l'espace à une dimension. Il faut en effet une deuxième dimension pour pouvoir les faire se coïncider.

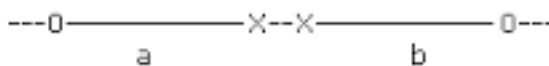


FIGURE 4.5 – La droite de Wittgenstein

L'on peut donc résumer l'objection de Wittgenstein en disant que si l'espace était non euclidien, il se pourrait que les énantiomères soient parfaitement identiques en soi, mais qu'ils nous paraissent différents en raison de notre intuition spatiale limitée. Ce ne serait donc qu'une illusion phénoménale. Cela dit, ce genre de stratégie semble tout de même aller dans le sens de Kant. Ce dernier, en effet, n'a jamais prétendu connaître les noumènes (ou choses en soi). Le seul fait d'admettre que des énantiomères puissent apparaître dans notre intuition spatiale semble attester de l'existence d'une telle intuition. Il reste donc possible d'affirmer que les objets nous sont donnés dans le cadre d'une forme *a priori* de la sensibilité qui en modifie les apparences. Si Kant perd du terrain en ce qu'il doit admettre que l'espace en-soi puisse exister conjointement à notre intuition, il conserve pour lui l'existence d'une forme *a priori* de la sensibilité qui puisse répondre d'une géométrie distincte de cet espace.

4.4.2 L'Espace est-il une construction ?

Un peu dans l'optique de Poincaré, l'on peut aussi se demander si notre intuition, loin d'être une structure *a priori* de la sensibilité, ne serait pas une construction culturelle ou empirique ? Dans notre monde à nous, les corps semblent être assez bien décrits à l'aide de la géométrie euclidienne, qui s'est silencieusement imposée dans nos descriptions mathématiques du monde. Mais si nous avions pu grandir dans un monde hyperbolique, n'est-il pas raisonnable de penser que notre intuition de l'espace eût été différente ? La principale réponse de Kant à ce genre de scepticisme consiste à dire la chose suivante : rejeter l'idée que l'espace est une intuition *a priori* (ou toute autre chose *a priori*) c'est se priver de la possibilité d'expliquer l'existence de propositions mathématiques *a priori*. C'est donc renoncer à fonder des vérités universelles. Si en effet l'espace procède d'une expérience, que cette expérience soit de l'ordre des sens ou qu'elle soit une convention de langage, cela signifie que tout ce que l'on connaît de lui, on ne le connaît que de façon *a posteriori*. L'on est donc mis en demeure d'expliquer d'où vient la puissance du raisonnement mathématique, et qu'est-ce qui lui donne plus de stabilité et de certitude, du moins en apparence, que les autres sciences

empiriques ? Les deux positions sont défendables, selon que l'on accorde plus d'importance à l'universalité des mathématiques ou à la relativité de nos représentations culturelles.

Notons que, pour Kant, l'espace n'est pas non plus un concept *a priori*, car le concept peut certes se référer à une infinité de choses possibles, comme le concept du triangle peut s'appliquer à une infinité de triangles possibles, mais aucun concept ne peut être lui-même infiniment divisible. Le concept d'une droite, par exemple, n'est pas infiniment divisible. Si l'on divise une droite, l'on obtient deux demi-droites, et non plus une droite. En revanche, la notion d'espace dans laquelle baigne le concept de la droite, pour sa part, est infiniment divisible. C'est pourquoi il ne peut pas voir dans l'espace un concept de l'Entendement, à la façon des idéalistes, et moins encore une construction opérée par l'esprit à partir d'expériences sensibles. Il s'agit au contraire d'une intuition *a priori* qui structure nos concepts. C'est, en d'autres termes, une condition de possibilité de la pensée elle-même. Il ne semble pas que les géométries non euclidiennes modifient quoi que ce soit à cet argument. En revanche, cela pourrait être le cas des géométries discrètes. Mais entrer dans le détail de cette question nous éloignerait de notre objet principal.

4.4.3 Les Mathématiques sont-elles synthétiques *a priori* ?

L'objection la plus grave est sans doute celle-ci. Kant soutient que toutes les propositions des mathématiques pures sont des propositions synthétiques *a priori*. D'autre part, il pense que l'intuition pure de l'espace et du temps est la forme *a priori* qui rend ces jugements possibles. Si notre intuition est, comme elle semble l'être, limitée à la géométrie euclidienne, *comment peut-on expliquer que des propositions mathématiques valides aient été conçues concernant des géométries non euclidiennes* ? Il faut soit (a) que ces propositions ne soient pas synthétiques *a priori*, soit (b) que notre intuition de l'espace englobe d'une certaine manière les géométries non euclidiennes.

Dans le cas (a), que deviendrait le statut des propositions mathématiques ? Une première option consiste à dire que les géométries elliptiques et hyperboliques sont le fruit de jugements analytiques, tandis que la géométrie euclidienne resterait synthétique *a priori*. Le problème consiste alors à justifier cette différence. Y a-t-il réellement quelques distinctions remarquables eu égard à leurs méthodes de raisonnement ? Cette méthode est, dans les deux cas, de nature hypothético-déductive. Ce sont les mêmes règles d'inférence, les mêmes symboles et — comble de l'ironie ! — quasi les mêmes hypothèses de départ qui guident la pensée dans un cas comme dans l'autre. La seule différence repose sur un unique postulat.

Mais peut-être dira-t-on que la différence pertinente est de l'ordre du contenu ? Les propositions de la géométrie euclidienne, dira-t-on, se rapportent à des intuitions *a priori*, tandis que celles des géométries non eu-

clidiennes ont perdu pied avec le contenu sensible. Cette remarque est tout à fait exacte. Mais si l'on prend au sérieux l'enseignement de Kant, il faudra conclure avec lui que les assertions des géométries non euclidiennes, en s'étendant hors de l'intuition, ne sont plus que des spéculations vides. Leur sort devrait être similaire à celui des antinomies de la raison : à plus ou moins brève échéance, l'on devrait voir apparaître des contradictions. Or il est démontré que ces théories sont vraies, ou pour être exact, qu'elles ne peuvent pas être plus fausses que la géométrie euclidienne. En interprétant chacun des termes d'une géométrie non euclidienne avec les termes de la géométrie euclidienne, il est en effet possible de montrer que si cette dernière est consistante, ce qui a été prouvé en 1951 par Tarski, alors la première l'est également. De plus, les progrès réalisés dans le domaine depuis plus de cent cinquante ans laissent à penser que, contrairement à la métaphysique spéculative, les géométries non euclidiennes représentent un ensemble de connaissances stables qui, bien loin de se réfuter au cours des décennies, n'ont eu de cesse de s'accroître et de former, progressivement, un royaume prospère.

Une alternative consisterait à voir les propositions des géométries non euclidiennes comme des jugements synthétiques *a posteriori*, tout en maintenant les jugements synthétiques de la géométrie euclidienne dans le domaine de l'apriorité. Mais cette position semble encore plus difficile à tenir. Elle signifierait que les résultats de la géométrie elliptique, qu'il est si difficile de se représenter à l'intuition, seraient le fruit d'expériences réalisées dans notre espace physique ? Or cet espace, à notre échelle d'observation directe, est bien plus proche de la géométrie euclidienne que de toute autre forme de géométrie. Comment donc un espace euclidien peut-il servir de support à l'expérience d'un espace non euclidien ? Sans évoquer le fait que cela semble en contradiction avec l'histoire de la science. Tout au long des deux derniers millénaires, les mathématiciens n'ont pas cessé de réfuter les géométries non euclidiennes. Il a fallu développer l'analyse logique pour que, peu à peu, celles-ci s'imposent à la communauté scientifique comme une possibilité. C'est seulement après ce progrès théorique que la physique, empruntant comme souvent ses outils à la reine des sciences, a transposé ces géométries nouvelles à la cosmologie. Notre réalité empirique, loin d'avoir contribué à leur développement, semble au contraire avoir été son principal obstacle, du moins à ses débuts. Et comme si tous ces arguments n'étaient pas suffisants, il reste encore à se demander, comme nous l'avons fait dans le paragraphe précédent, quelle est la différence de méthode ou de contenu entre les géométries euclidiennes et les géométries non euclidiennes qui puisse justifier une différence de traitement épistémologique ?

Puisqu'il paraît difficile d'attribuer aux propositions de la géométrie euclidienne un autre statut que celui conféré aux géométries non euclidiennes, il reste pour unique possibilité de réviser la croyance comme quoi *toutes les propositions mathématiques sont des jugements synthétiques a priori*. Ceci

ouvre à de multiples alternatives. Premièrement, l'on peut supposer qu'elles ne sont finalement que des jugements analytiques — et que Kant avait tort d'avoir voulu sortir de là. C'est la voie qu'adopteront les logiciens lorsqu'ils tenteront de faire découler les mathématiques d'un système formel approprié. Une autre solution consisterait à les voir comme des jugements synthétiques *a posteriori*. Cela renvoie alors aux conceptions philosophiques de Poincaré (pour le conventionnalisme) et de Lobatchevski (pour le constructivisme empirique). Le prix à payer, comme nous l'avons fait remarquer plus haut, c'est une perte d'universalité. Il devient difficile de défendre l'apriorité des mathématiques ; en tout cas pas dans un sens aussi pur et rigoureux que Kant le voulait⁵⁷

Le cas (b) semble représenter également une solution envisageable. Il consiste à substituer une nouvelle forme d'intuition pure à celle de l'espace et du temps. Cette intuition ne peut plus avoir la forme d'une simple évidence, mais doit s'accommoder du fait que, dans le cas des géométries euclidiennes, les objets ne nous sont donnés à voir que dans des interprétations intuitives. La géométrie hyperbolique, par exemple, peut être conçue à l'aide du disque de Poincaré ; la géométrie elliptique peut être vue sur une sphère dont on identifie les points antipodaux. Tous ces modèles simplifiés sont des projections sur un plan accessible à l'intelligence humaine. L'intuition n'est donc plus cet horizon de donation clair et distinct, où les choses se présentaient identiquement à elles-mêmes. Elle est, dans le meilleur des cas, un espace où les objets sont donnés à voir sous une certaine facette. Cette idée n'est pas sans évoquer plusieurs conceptions philosophiques du $XX^{\text{ème}}$ siècle. Je pense d'une part au principe de complémentarité de Bohr ou aux *abschattungen* de Husserl.

Notons, pour finir cette discussion, que si l'intuition pure du temps a joui d'une position plus confortable que celle de l'espace tout au long du $XIX^{\text{ème}}$ siècle, elle ne fut pas épargnée à l'entrée dans le $XX^{\text{ème}}$. Les progrès de la physique ont montré que le temps n'était pas indépendant du système de référence. En 1905, Einstein publie un article, devenu célèbre, démontrant que le temps se dilate en fonction de la vitesse du système de référence galiléen conformément aux transformations de Lorentz — connues par ailleurs. Dix ans plus tard, il construit un modèle capable de généraliser ces résultats à tous les systèmes de référence. Ces importants travaux, réalisés avec la coopération de Minkowski, sont l'occasion d'introduire une représentation de l'espace-temps à 4 dimensions, qui finit de bouleverser (1) notre intuition du temps et (2) la croyance en une intuition du temps indépendante de celle de l'espace. Cela dit, ces progrès ne signifient pas la mort de l'intuitionnisme en tant que tel, que du contraire ! Kant continuera d'influencer des générations

57. L'empirisme doit aussi rendre compte du fait que nous possédons la faculté de structurer les données de notre expérience. Comment se peut-il que cette capacité, qui rend l'expérience possible, nous vienne de l'expérience elle-même ?

entières de mathématiciens, et au premier plan Hilbert, dont nous allons voir qu'il a donné à l'axiomatique une nouvelle conception épistémologique.

— Ce qu'il faut retenir de ce chapitre —

Kant présupposait l'existence de deux structures innées chez l'être humain : *l'intuition du temps* et *l'intuition de l'espace*. Il est commode d'utiliser ces intuitions pour expliquer les différences de jugement entre l'esprit humain et nos algorithmes contemporains.

L'histoire des géométries non euclidiennes semble toutefois contredire l'idée que ces intuitions soient aussi nécessaires qu'il le paraisse à l'exercice du jugement. On peut se demander, en particulier, s'il existe bien et bien un *jugement synthétique a priori*.

C'est pourquoi il semble judicieux de se mettre en quête des structures les plus fondamentales de l'esprit humain. Si ce n'est pas l'intuition du temps et de l'espace, de quoi peut-il s'agir ?

Je tenterai d'y répondre dans la suite de cet essai. Mais auparavant, il faut comprendre comment l'abandon de la théorie kantienne a donné naissance au développement théorique des premiers ordinateurs. C'est l'objet de la prochaine partie.

Chapitre 5

La Rationalité numérique : Naissance de l'ordinateur

J'appelle ici "calcul" toute
solution qui représente le
raisonnement, même si elle n'a
aucun rapport avec les nombres.

Gottfried Wilhelm Leibniz

Le bouleversement provoqué par les géométries non euclidiennes ne vient pas tant de l'abandon du cinquième postulat d'Euclide que de ce constat inattendu : les vérités mathématiques peuvent déborder largement du cadre de nos intuitions. Depuis l'Antiquité déjà, on savait que le plus court chemin entre deux points du globe terrestre était la géodésique. Il était aussi connu que la somme des angles d'un triangle dessiné sur une sphère était supérieure à 180 degrés. De tels résultats ne choquaient pas notre expérience usuelle de l'espace ni nos représentations admises.

Or au *XIX^{ème}* siècle, l'on découvre soudain que le raisonnement pur détient une puissance insoupçonnée. En changeant les prémisses de la théorie d'Euclide, pourtant bien établie, et en explorant les conséquences logiques sans faire intervenir l'intuition humaine, ce sont des mondes nouveaux, étranges et hallucinés, qui se dévoilent soudain au regard humain. La conviction s'impose peu à peu que le formalisme dépasse le contenu intuitif. La vérité, loin de se tenir sagement dans la sphère des jugements synthétiques *a priori*, tombe sous le joug du jugement analytique. Elle découle tout entière des déductions logiques produites à partir d'un petit nombre d'axiomes. Ces derniers peuvent être choisis de différentes manières, selon les besoins pragmatiques et les conventions établies, sans recourir à la perception sensible, à l'intuition ou à l'entendement du mathématicien.

Ces interrogations sont fâcheuses eu égard au projet philosophique que j'ai défendu jusqu'à présent. Ma thèse consistait à dire que le jugement syn-

thétique *a priori* pourrait être l'élément qui fait encore et toujours défaut à nos algorithmes. Cette opinion jouissait des meilleurs auspices. Elle semblait recevoir l'aval de la théorie philosophique du jugement (de Platon à Kant) et de la pratique algorithmique (les exercices ARC). Mais un nuage est venu obscurcir l'espoir d'un triomphe facile : en parcourant l'histoire des géométries non euclidiennes, il est apparu que les notions kantienne d'intuition pure de l'espace et du temps pouvaient être prises à défaut.

Est-ce à dire que le jugement synthétique *a priori* doit être abandonné ? Doit-on le voir comme un reliquat de la pensée humaine, un peu à la façon du liquide calorique, de l'éther, du fluide boréal, du magnétisme animal et de toutes ces fausses théories scientifiques ? Ce serait un brin radical. Il semble toujours exact de dire que notre représentation du monde s'inscrit dans une géométrie euclidienne et que notre représentation du temps épouse la ligne du temps newtonienne. Il faut en outre reconnaître à Kant le mérite d'avoir défendu, de façon totalement visionnaire, l'idée que ces intuitions sont en nous, qu'elles sont des structures de la représentation, et qu'elles ne correspondent pas forcément à la réalité du monde.

En dépit des graves nuages qui pèsent sur le jugement synthétique *a priori*, il semble toujours prometteur de penser qu'il y a bien une structure, dans l'esprit humain, qui nous fait voir l'information différemment de la machine. N'ai-je pas montré, au travers des illusions d'optique, que l'humain se représente les données de façon holistique ? N'ai-je pas suggéré que la machine a du mal à réaliser des abstractions ? C'est pourquoi la tâche essentielle qui m'incombe, si je ne veux pas signer l'arrêt de mort de ma démarche, consistera à donner au jugement synthétique *a priori* une nouvelle définition. Cette définition devra éviter les deux écueils qui consistent, d'une part, (1) à supposer l'existence d'une intuition euclidienne de l'espace-temps et (2) à réduire les processus de pensée à de simples calculs logiques.

Cette révision en règle du concept de jugement synthétique *a priori* exigera, dans un premier temps, de comprendre plus précisément les paradigmes de l'informatique théorique ayant donné naissance au rêve que « tout est calculable ». C'est en déterminant les limitations de ces paradigmes et en comprenant sur quels genres de questions ils échouent à apporter une réponse satisfaisante, que j'entends trouver des ingrédients pour construire une nouvelle théorie du jugement humain. Cette stratégie permettra, du reste, de mieux comprendre le fonctionnement de la Machine, entendue comme un mode spécifique de rapport à l'information.

Dans ce chapitre, j'aimerais reprendre l'étude du jugement humain là où les théories non euclidiennes l'avaient laissée : sur le triomphe du pur raisonnement face aux intuitions spatio-temporelles. Partant de cette friche désolée, il s'agira de montrer que l'abandon progressif des formes *a priori* de Kant a conduit la philosophie à représenter le monde uniquement au moyen de propositions calculables. À toute question scientifique, précise et rigoureuse, doit pouvoir être apportée une réponse tout aussi exacte. Le processus

qui va de la question à la réponse est doté de règles claires, automatisables, similaires en cela à la division euclidienne. Si vous voulez calculer la division de dix par trois, la procédure à suivre est prédéterminée. Vous n'avez rien à faire qu'à suivre les étapes. De la même manière, l'on souhaiterait de la science qu'elle puisse être automatisée de sorte à ne pas laisser place à l'erreur. Ce rêve, cet idéal, est popularisé par le mathématicien David Hilbert. Il donnera naissance à la toute première forme d'intelligence artificielle : un être humain munis d'un bloc de feuilles et d'un crayon, qui applique des règles de calcul, qui dérive des équations, qui pond des symboles sur le blanc de la page, sans jamais faire intervenir son jugement propre.

5.1 Le Rêve d'un Calculateur universel

Le 6 août 1900, le Palais des Congrès est en pleine effervescence. On peut y voir des savants en chapeaux melons, serviette sous le bras et pipe en poche, défiler sous les arcades. D'aucuns parlent de géométrie, d'autres d'algèbre et d'arithmétique. L'un d'entre eux, au chapeau plus imposant que ses congénères, prend l'initiative d'ouvrir la séance. David Hilbert introduit, avec Henri Poincaré, le Congrès international des mathématiques. Lors de son discours en séance plénière, qui aura lieu à la Sorbonne quelque temps plus tard, il fait la liste de vingt-trois problèmes ouverts en mathématiques.

L'un d'entre eux, le deuxième problème, est particulièrement intéressant pour notre propos. Il consiste à se demander si on peut axiomatiser l'arithmétique de façon complète et cohérente. L'arithmétique, c'est l'ensemble des entiers positifs $0, 1, 2, \dots$ dotés des opérations d'addition et de soustraction. Cette arithmétique est dite complète s'il est possible de donner une liste finie d'axiomes (c'est-à-dire de propositions considérées vraies en elles-mêmes) sur la base desquels toutes les vérités arithmétiques puissent être dérivées, comme " $5 + 7 = 12$ " ou "il est possible d'associer tout nombre pair à un et un seul nombre impair". De plus, cette arithmétique est dite cohérente si aucune contradiction ne peut résulter de cette liste d'axiomes.

Pour illustrer le problème, il peut être utile de préciser que la géométrie euclidienne est dotée d'une axiomatique complète et cohérente. Toutes les vérités qu'il est possible d'énoncer à propos des droites, des points, des triangles et des cercles dans le plan sont en effet démontrables sur base d'un petit ensemble d'axiomes, appelés les axiomes d'Euclide :

- **Axiome 1** : Il existe toujours une droite qui passe par deux points du plan.
- **Axiome 2** : Tout segment peut être prolongé en une droite infinie.
- **Axiome 3** : À partir d'un segment, il existe un cercle dont le centre est un des points du segment et dont le rayon est la longueur du segment.

- **Axiome 4 :** Tous les angles droits sont égaux entre eux.
- **Axiome 5 :** Étant donné un point et une droite ne passant pas par ce point, il existe une seule droite passant par ce point et parallèle à la première.

Or il se trouve que, à l'époque de Hilbert, personne n'est parvenu à donner une liste d'axiomes analogues pour l'arithmétique. Hilbert imagine alors la fiction d'un mathématicien disposant d'un crayon et d'un bloc de feuilles de papier en quantité illimitée. Un tel être humain parviendra-t-il à écrire, sous la forme d'une combinaison finie de signes, une preuve à tous résultats arithmétiques vrais, et rien que ceux-là ?

Plutôt que d'insister sur l'existence des entités abstraites, telles que les nombres, les opérations ou les groupes algébriques, Hilbert considère le *signe* comme la seule définition admissible d'un objet mathématique. Ce signe est tout ce qu'il y a de plus palpable : il existe sur un support papier, suit des règles de syntaxe très strictes et toutes les manipulations dont il peut faire l'objet sont visibles, traçables, vérifiables. Le problème de l'axiomatisation de l'arithmétique revient alors à trouver l'ensemble des règles, c'est-à-dire l'algorithme, qui régit l'arithmétique. Il faut en particulier que :

- L'algorithme consiste en un ensemble fini d'instructions simples, elles-mêmes décrites avec un nombre limité de signes ;
- L'algorithme produise le résultat en un nombre fini d'étapes ;
- L'algorithme ne requiert pas plus d'intelligence, de la part de l'exécuteur, que celle qui est requise pour suivre les instructions.

On peut dire que Hilbert pose les jalons du concept de calculabilité, tel qu'il sera repris ultérieurement par les mathématiciens Church et Turing dans leur thèse éponyme (la thèse de Church-Turing). Il n'est d'ailleurs pas anodin que le terme *computer*, en anglais, désignera pendant longtemps un homme capable de produire des calculs sur le papier, et à qui l'on confiait la tâche de calculer des coefficients de résistance, des tractions ou des trajectoires balistiques. Ce n'est qu'après la mise aux points des premiers ordinateurs que le terme changera progressivement de signification pour devenir celui que l'on connaît aujourd'hui.

Au discours de Hilbert succédera, en l'espace de quelques décennies, deux types de résultats démontrant, chacun dans des cadres différents, qu'il est impossible d'exprimer toutes les vérités mathématiques avec des règles, un crayon et du papier. Les deux cadres en question regroupent, d'une part, les théories de la démontrabilité et, d'autre part, les théories de la calculabilité. Si les premières représentent le raisonnement à l'aide de systèmes d'axiomes formels à la mode d'Euclide, les secondes, quant à elles, stipulent qu'un raisonnement est un algorithme à la mode de Hilbert.

Définition 2 (Paradigme de la démontrabilité). *Tout calcul peut s'exprimer comme une déduction logique au sein d'un système formel. Un système formel est composé :*

- **D'axiomes** : Des propositions jugées vraies en elles-mêmes.
- **De règles de raisonnement** : Des lois déterminant comment inférer une proposition à partir d'une autre.

Définition 3 (Paradigme de la calculabilité). *Tout calcul peut être réalisé par une machine de Turing. Cette dernière est composée :*

- **D'un ruban** : Un bout de papier décomposé en une infinité de cases. Chaque case peut comporter un symbole ou rester vierge.
- **D'une tête de lecture** : Un appareil mécanique capable de se déplacer sur le ruban, de lire les symboles, de les effacer et d'en écrire de nouveaux.
- **D'un programme** : Un ensemble d'instructions disant à la tête de lecture ce qu'elle doit faire. Par exemple : "Si le symbole "a" est inscrit sur la première case, effacer le symbole écrit en troisième case."

Un "calcul" consiste à adresser une question à la machine en l'écrivant sur le ruban et à lire la réponse que cette dernière produit, après un temps fini, sur ce même ruban.

Ces deux paradigmes épousent en quelque sorte le rêve des caractéristiques universelles de Leibniz, lorsque ce dernier s'écriait, à propos de n'importe quel problème philosophique ou théologique : « Calculons ! ». Deux importants résultats mathématiques démontrent cependant, pour chacun de ces paradigmes, qu'il existe certains problèmes incalculables. Il est même permis de dire que les seuls problèmes auxquels l'on puisse apporter une solution par la voie du calcul ne sont rien que des points isolés dans l'espace infini des propositions mathématiques possibles. Mais en raison de la finitude de l'esprit humain, seules les questions solubles les plus évidentes ont fait l'objet d'une attention soutenue dans l'histoire des mathématiques ; si bien que la part immergée de l'iceberg est longtemps passée pour inexistante.

Voici un succédané des deux résultats en questions. Je vais les prendre un par un et les développer dans les prochains chapitres.

Théorème 1 (Théorèmes d'incomplétude de Gödel). *Tout système d'axiomes A suffisamment fort pour exprimer les propositions de l'arithmétique possède au moins une proposition indémontrable (et dont la négation est elle aussi indémontrable). C'est le cas en particulier de la proposition stipulant : « A est non-contradictoire ».*

Théorème 2 (Problème de l'arrêt). *Il est impossible de construire une machine de Turing capable de dire, de toute autre machine de Turing, si elle*

va donner une réponse à un problème donné en un temps fini ; ou si elle va tourner sans jamais s'arrêter.

5.2 Les Théories de la démontrabilité

Gottlob Frege est un logicien tout ce qu'il y a de plus caricatural. Il est précis, austère, obstiné — et d'une intégrité intellectuelle radicale. Son caractère l'aura poussé, avec une évidence dont la nature a le secret, vers le projet de trouver les axiomes de l'arithmétique. Ce projet le caresse assez tôt dans sa carrière. Il s'imagine produire, pour l'addition, la soustraction et la multiplication des nombres entiers, ce qu'Euclide avait fait, des milliers d'années avant lui, pour les carrés, les cercles et les triangles. Cela lui permettrait d'apporter une réponse concrète, opérationnelle, au rêve d'automatisation de la pensée mathématique formulé par David Hilbert.

Au départ de ce projet philosophique, Frege commence par proposer une redéfinition des concepts touchant au « synthétique » et à « l'analytique ». Pour déterminer la nature d'une proposition analytique, affirme Frege, il faut en analyser les étapes de la démonstration, en partant des conclusions et en remontant jusqu'aux axiomes. S'il s'avère, à l'issue de cet examen, que la proposition ne dépend de rien d'autre que de lois logiques, elle est qualifiée *d'analytique*. Mais si au moins une loi issue des autres sciences est incorporée, alors la proposition est *synthétique*.

De plus, la proposition est dite *a priori* si elle ne repose que sur des lois fondamentales, comme par exemple les principes de la mécanique newtonienne, et elle est *a posteriori* si elle incorpore des données empiriques. Les propositions analytiques, selon cette définition, sont donc toutes *a priori*. Mais ce ne sont pas que des tautologies, loin de là. Le point de vue de Frege, sa thèse fondamentale, consiste à dire, contra Kant, que toutes les propositions de l'arithmétique, telle que « $7 + 5 = 12$ », sont analytiques¹. Cette position pose les jalons du logicisme. Elle s'oppose aussi bien à Kant, pour qui les propositions mathématiques sont synthétiques *a priori*, qu'à des auteurs anglophones comme Henri Poincaré ou John Stuart Mill, pour qui elles sont synthétiques *a posteriori*.

Frege entreprend de prouver cette thèse logiciste dans son maître ouvrage, intitulé *Les Fondements de l'arithmétique*. À cette fin, il commence par définir un ensemble de lois de base et de règles d'inférence, ainsi que des définitions leur associant des dénominations. Ces définitions ne créent rien en tant que telles, elles ne font que donner un nom aux termes impliqués. Comme il le dit lui-même : « Les définitions ne créent pas réellement quoi que ce soit, à mon avis, et elles ne le doivent pas ; elles introduisent simplement des notations abrégées (noms), dont on pourrait se passer si la longueur

1. BÉLANGER MATHIEU, *La méthode axiomatique durant la crise des fondements* in *La crise des fondements : quelle crise ?*, p. 46.

n'entraînait pas alors des difficultés extérieures insurmontables.^[2] » Quant au statut des lois de base ? Leur rôle est fondationnel. Ce sont des axiomes « évidents par eux-mêmes ».

Ce qu'il faut entendre exactement dans cette « l'auto-évidence », qui est le fer de lance de la pensée de Frege, fait l'objet de toute une littérature philosophique. Selon Robin Jeshion, de l'université de Californie du Sud, le terme comporterait deux sens :

Plus spécifiquement, mon opinion est que Frege a deux notions d'auto-évidences. La première notion consiste à dire de la vérité qu'elle a des fondations solides et qu'elle ne dépend pas d'une autre vérité. La seconde notion consiste à dire de la vérité qu'elle peut être saisie de façon claire, rationnelle et a priori^[3]

La loi de base devrait être « évidente » dans les deux sens du terme avant de pouvoir être admise. Le premier sens semble traduire un réquisit pragmatique : le fait que l'axiome ne doit dépendre d'aucun autre axiome. Le second semble renvoyer à un critère épistémique : celui de la clarté, en un sens quasi cartésien. Frege ne suppose donc pas l'existence d'une intuition dans laquelle les évidences seraient données à voir. L'évidence qu'il a en vue est plutôt celle d'un acte de langage.

Toujours est-il que Frege, muni de cette auto-évidence, entreprend de chercher les axiomes au fondement de l'arithmétique. Au terme de plusieurs années de recherche, il consigne ses conclusions dans un livre voué à un destin tragique.

5.2.1 Le Paradoxe du barbier

Les *Fondements de l'arithmétique* de Frege sont publiés en 1884, mais le projet dont il est porteur est inopinément avorté lorsque Russell identifie le paradoxe qui porte son nom. Dans la lettre qu'il envoie à Frege en 1902, il explique que la façon dont les objets mathématiques sont définis semble conduire à des contradictions. Il pointe du doigt, tout particulièrement, le concept de « classe » :

« De même, il n'existe pas de classe (en tant que totalité) des classes qui, prises chacune comme une totalité, ne se contiennent pas elles-mêmes. J'en conclus que, dans certaines circonstances, une collection définie ne forme pas une totalité.^[4] »

2. FREGE, *Les Fondements de l'arithmétique*, p. vi. La traduction est de moi.

3. Voir JESHION ^[93], *Frege's Notions of Self-Evidence*. La traduction est de moi.

4. Lettre de Russell du 16 juin 1902. Cité depuis LINSKY, *Ernst Schroeder and Zermelo's Anticipation of Russell's Paradox* in *La crise des fondements : quelle crise ?*, p. 8., p.

9. La traduction est de moi.

Frege lui écrira six jours plus tard « Il semblerait que ma règle V (§20,p.36) soit fausse, et que mes explications en §31 sont insuffisantes pour assurer que mes combinaisons de signes aient une signification dans tous les cas⁵ »

Il est possible de donner à ce paradoxe de Russell une illustration plus intuitive, connue sous le nom du *paradoxe du Barbier*. Pour ce faire, imaginons un barbier dont la boutique fait, chaque jour, l'objet de files interminables. Afin d'écourter le temps d'attente, la commune fait voter la résolution suivante : « Dorénavant, le barbier devra raser tous ceux qui ne se rasent pas eux-mêmes et personne d'autre que ceux-là. » Bien que raisonnable à première vue, cette nouvelle loi va avoir, pour notre brave homme, une conséquence imprévisible. Le septième jour, en effet, se présente dans le village un logicien connu sous le nom de Russell. Pipe à la bouche et chapeau sous le bras, le nouveau venu interroge à la façon de Sherlock Holmes : « Dites-moi, mon cher barbier, au vu des lois de votre village, pouvez-vous vous raser la barbe vous-même ? »

La question est épineuse. Car si le barbier répond oui, il enfreint la loi, puisqu'il lui est interdit d'offrir ses services à ceux qui se rasent eux-mêmes. Mais s'il répond non, il enfreint également la loi, puisqu'il est censé raser tous ceux qui, comme lui, ne se rasent pas eux-mêmes. L'histoire ne raconte pas ce qu'il advint du barbier, mais l'on prétend que, devant le tribunal, il aurait plaidé coupable⁶.

Une deuxième illustration du paradoxe, plus succincte encore, nous vient de Martin Gardner^{footnote}On la lui attribue généralement, mais je n'ai pas réussi à retrouver la référence exacte.. Elle consiste à imaginer une encyclopédie qui citerait les noms de toutes les encyclopédies ne se citant pas déjà elles-mêmes. Le paradoxe surgit lorsque l'on se pose la question suivante : l'encyclopédie de Gardner doit-elle s'autoréférencer ?

Ce qui peut sembler un innocent paradoxe plonge le logicisme dans un état de profonde perplexité. La seule manière, pour Frege, de contourner l'antinomie, c'est de modifier la règle V de façon à éviter qu'une classe puisse se contenir elle-même. De manière générale, tous les logiciens postérieurs à cette lettre de 1902 auront à répondre de cette difficulté. Comme le fait remarquer Poincaré dans *Les Mathématiques et la logique*, le cercle vicieux vient de ce que l'on suppose qu'une collection d'objets est capable de contenir des éléments qui, à leur tour, ne peuvent être définis qu'au moyen de la collection prise comme un tout⁷.

Par exemple, une proposition P énonçant que « Toutes les propositions sont soit vraies, soit fausses » est problématique. En effet, à quelles propositions se rapporte P ? Sans doute, dira-t-on, à « toutes les propositions » existantes. Or comment définir cet ensemble de « toutes les propositions »

5. Réponse de Frege à Russell, 22 juin 1902. Cité depuis *ibidem*, p. 8.

6. Une variante rapporte qu'il se serait laissé pousser la barbe jusqu'à ce qu'elle soit assez longue pour étrangler tous les logiciens du village...

7. POINCARÉ, *Les mathématiques et la logique*, p. 307.

intervenant dans P si la seule énonciation de P crée déjà une nouvelle proposition ? Il faut donc abandonner les termes du type « toutes les propositions », et plus généralement, les termes se référant à tous les membres d'un ensemble⁸.

5.2.2 Le Paradoxe de Cantor

Un autre paradoxe d'importance majeure, dans l'histoire des mathématiques, est celui démontré par Cantor en 1899. Bien qu'il soit chronologiquement antérieur au paradoxe de Russell, il est commode, d'un point de vue strictement logique, de le présenter en second lieu. Au lieu de s'intéresser à l'ensemble des ensembles qui ne se contiennent pas eux-mêmes, ce paradoxe se penche sur l'ensemble de tous les ensembles — tout court. Il y a donc bien, dans un cas comme dans l'autre, un procédé non constructif : on pose l'existence d'un ensemble infini sans en expliciter le contenu. Si toutefois le paradoxe de Russell présente une contradiction évidente, on ne voit pas nécessairement quel problème logique découlerait du fait de parler de l'ensemble de tous les ensembles.

Je vais tenter d'expliquer le paradoxe. En guise de première observation, il est important de constater que l'ensemble de tous les ensembles, noté ENS , est infini et autoréférentiel. Infini, car il existe une infinité d'ensembles possibles. Autoréférentiel, car par définition, ENS doit contenir tous les ensembles, y compris lui-même. Il existe donc un élément $x \in ENS$ tel que $x = ENS$. Cela veut dire que l'ensemble ENS peut être conçu comme un grand sac qui contiendrait, à l'intérieur, un sac avec lui-même. On sent qu'il y a quelque chose, dans la notion d'appartenance, qui heurte le sens commun. Comment peut-on s'appartenir soi-même ? Il n'est toutefois pas rare d'avoir des propriétés contre-intuitives, lorsque l'on manipule des objets infinis.

Pour montrer rigoureusement l'existence d'un paradoxe, Cantor va procéder à un argument connu sous le nom « d'argument de la diagonale. » Je vais montrer comment cet argument peut être utilisé pour conclure que, dans un univers qui ne contiendrait que des entiers naturels $1, 2, 3, \dots$, l'ensemble de tous les ensembles d'entiers naturels ne peut pas se contenir lui-même. Je vais ensuite généraliser le raisonnement à l'ensemble ENS de tous les ensembles. Un lecteur désireux d'avoir plus de détails est prié de se référer à l'annexe 4, qui expose la théorie de Cantor en long et en large.

8. Les termes se référant à tous les membres d'un ensemble ne sont pas toujours problématiques, et ils le sont même "rarement". On peut par exemple dire : « tous les nombres naturels sont supérieurs ou égal à zéro. » et il n'y a là rien de paradoxal. Mais une logique permettant de quantifier tous les membres d'un ensemble infini peut conduire à des paradoxes, comme le paradoxe de Russell. Une manière d'éviter ces paradoxes consiste à ajouter des contraintes sur les termes généraux, ce qu'on appelle, d'un point de vue technique, la quantification infinie.

L'argument de la diagonale de Cantor

L'argument de la diagonale peut être utilisé pour montrer que l'ensemble $\mathbb{N}$ des entiers naturels n'est pas assez grand pour contenir l'ensemble des sous-ensembles d'entiers naturels, noté $\mathbb{P}(\mathbb{N})$. Pour rappel, un sous-ensemble des entiers naturels est n'importe quel ensemble comportant zéro, un ou plusieurs naturels. Par exemple, $\{1, 5, 7\}$ est un sous-ensemble de $\mathbb{N}$, mais aussi l'ensemble des nombres pairs $\{2, 4, 6, \dots\}$, l'ensemble vide $\{\}$ et $\mathbb{N}$ lui-même.

Pour arriver à ce résultat, supposons qu'il existe un livre L doté d'une infinité dénombrable de pages et dont chaque page comporterait un ensemble de naturels distincts. La page numéro 1 présenterait par exemple l'ensemble $\{1, \{2, 3\}\}$, la page 2 comporterait l'ensemble $\{\{7\}\}$, etc. Cantor démontre, dans un formalisme mathématique bien à lui, que si on admet l'existence de ce livre, alors on arrive à une contradiction. Cette stratégie de preuve, qui consiste à supposer l'existence de ce qu'on souhaite réfuter de façon à en montrer l'absurdité, est la *reductio ad absurdum*. On a déjà été confronté à ce genre de stratégie au moment de parler des géométries non-euclidiennes (voir la section 4.2).

Supposons donc que le livre de Cantor existe. Appelons "nombre ordinaire" un entier n qui n'est pas représenté sur la page portant le numéro n de ce livre. Dans notre exemple, le nombre 2 est un nombre ordinaire, puisque l'ensemble représenté sur la page numéro 2, à savoir $\{\{7\}\}$, ne comporte pas le nombre 2. Le nombre 1, en revanche, n'est pas un nombre ordinaire dans notre exemple, puisque la page numéro 1 contient le nombre 1 dans l'ensemble qu'elle représente : $\{1, \{2, 3\}\}$.

Comme le livre de Cantor comporte tous les ensembles de naturels, il doit exister une page, portant le numéro n' , contenant l'ensemble de tous les nombres ordinaires, et aucun autre. L'ensemble représenté sur la page n' devra donc, entre autres, contenir le nombre 2, mais pas le nombre 1. Vous me suivez ? On pose dès la question suivante : n' est-il un nombre ordinaire ?

Supposons tout d'abord que n' soit un nombre ordinaire. Il doit donc être représenté sur la page qui contient tous les nombres ordinaires. Or cette page porte le numéro n' . Cela contredit la définition de nombre ordinaire. Supposons maintenant que n' ne soit pas un nombre ordinaire. Par définition, cela veut dire qu'il doit être représenté sur la page portant le numéro n' . Cependant, cette page ne comporte que les nombres ordinaires, et aucun autre. On en conclut que le livre de Cantor est paradoxal.

Si on veut éviter que les mathématiques se contredisent, il faut refuser l'existence du livre de Cantor. Qu'est-ce que cela implique-t-il d'un point de vue logique ? Qu'il est impossible d'énumérer tous les ensembles de nombres naturels en les consignant dans un livre infini. Autrement dit, il est impossible que le livre qui contient tous les nombres naturels contiennent également l'ensemble des sous-ensembles de naturels.

L'ensemble de tous les ensembles

L'argument de la diagonale de Cantor a démontré que l'ensemble des naturels $\mathbb{N}$ n'était pas assez grand pour contenir les sous-ensembles de naturels. Mais l'argument peut être généralisé à n'importe quel type d'ensembles. Par exemple, on peut l'utiliser pour dire que l'ensemble des réels n'est pas assez grand pour contenir en lui-même tous les sous-ensembles de réels, notés $\mathbb{P}(\mathbb{R})$. Ce dernier ensemble, à son tour, n'est pas assez grand pour contenir l'ensemble $\mathbb{P}(\mathbb{P}(\mathbb{R}))$ des sous-ensembles de $\mathbb{P}(\mathbb{R})$, etc. Il existe donc une hiérarchie croissante d'ensembles imbriqués les uns dans les autres. La question que l'on se pose consiste à se demander s'il existe une sorte de super ensemble *ENS* qui contiendrait tous les ensembles possibles, y compris lui-même ?

La réponse est négative. Et pour cause : l'argument de la diagonale de Cantor peut être utilisé sur *ENS* lui-même. Si l'on suppose que cet ensemble existe, il ne serait donc pas assez grand pour contenir l'ensemble $\mathbb{P}(\text{ENS})$ des sous-ensembles de *ENS*. Autrement dit, il est toujours possible d'imaginer un ensemble plus grand que le plus grand ensemble que vous vous représentez. Il s'ensuit que l'ensemble de tous les ensembles est tout aussi contradictoire, en fin de course, que l'ensemble des ensembles qui ne contiennent pas eux-mêmes — imaginé par Russell.

5.2.3 Contourner les paradoxes

Différentes tentatives sont menées en parallèle pour contourner le paradoxe de Cantor et le paradoxe du barbier. Une première solution consiste à associer chaque terme de la théorie à un certain type. Le type minimal correspond aux individus, on trouve ensuite le type des ensembles ou des prédicats d'individus, puis le type des ensembles d'ensembles ou des prédicats de prédicats, etc. Avec cette solution, un ensemble ne peut pas se contenir lui-même, mais il peut uniquement contenir des objets associés au type inférieur. La théorie des types prévoit aussi qu'il existe des propriétés qui peuvent s'appliquer à elles-mêmes, et d'autres qui ne le peuvent pas pour éviter de générer des paradoxes. Cette théorie des types a été inspirée des *Principia Mathematica* de Russell et Whitehead. Ces derniers présentaient cependant, en plus des types, une hiérarchie sur les ordres, ce qui la rendait difficile à manipuler.

Une autre approche, qui rencontrera le plus de succès historique, est la formalisation de Zermelo-Fraenkel, notée ZF. Cette théorie décrit ce qu'est un ensemble mathématique, de "taille" potentiellement infinie^[9]. Bien qu'il ne s'agisse pas d'une théorie de la logique mathématique, elle a ceci de particulier qu'elle peut être exprimée dans une logique du premier ordre. Elle

9. La taille d'un ensemble fini dépend du nombre d'éléments qu'il contient et est appelée la cardinalité. La taille d'un ensemble infini peut s'entendre en plusieurs sens. Les définitions rigoureuses nécessitent l'introduction des nombres transfinis. Voir à cet effet l'annexe 5.

permet en outre d'éviter adéquatement les paradoxes qui surviennent dans la démarche de fondement des mathématiques, et ce, à l'aide d'un ensemble d'axiomes assez intuitifs. L'on pourrait dire, en un certain sens, que la crise des fondements a été résolue, mais ce serait oublier, outre le caractère technique de cette crise, la question épistémologique plus fondamentale qu'elle soulève : quel est le statut des axiomes ? Sont-ils suffisants pour permettre d'asseoir *tous* les résultats mathématiques ? Sont-ils, d'autre part, suffisamment *évidents* ? En quel sens ?

Un certain nombre d'axiomes, comme l'axiome du choix (AC) ou l'hypothèse du continu (HC), sont indépendants des axiomes de base de la théorie ZF et peuvent lui être adjoints. On note ZFC et $ZF + HC$ (respectivement) les deux nouvelles théories obtenues^[10]. Ajouter ces axiomes rend vrais des théorèmes qui heurtent l'intuition. Ne pas le faire rend la théorie incapable de dériver certains résultats fondamentaux de l'algèbre et de l'analyse.

Je vais illustrer cette situation avec l'axiome du choix. Dans sa formulation la plus simple, cet axiome stipule que « tout ensemble S qui ne contient pas l'ensemble vide a une fonction de choix ». Cette fonction de choix associe, à tout x de S , l'un des éléments de x . Formellement, une fonction $f : S \rightarrow \cup S$ est une fonction de choix si $\forall x \in S, f(x) \in x$. En termes intuitifs, l'axiome du choix consiste donc à dire que, lorsqu'on est en présence d'un ensemble de boîtes de pralines non vides, il est possible de définir une fonction qui décidera à l'avance quelle praline choisir dans chacune des boîtes qui se présentera à nous.

De prime abord, cet axiome n'a pas de quoi effrayer un mathématicien. Et pourtant, la subtilité vient du fait qu'il reste d'application même en présence d'une boîte contenant un nombre infini de pralines. En fait, il est possible de montrer un certain nombre de résultats étonnants sur base de cet axiome, et notamment celui-ci : en utilisant l'axiome du choix, une sphère d'un volume V donné peut être décomposée en une infinité de parties, puis reconstituée pour former une sphère de volume arbitraire. Les conséquences de l'axiome sont donc clairement en contradiction avec notre intuition de l'espace^[11].

10. L'indépendance de l'axiome du choix a été montrée par Cohen en 1969. À cette époque, on savait déjà, grâce à un résultat de Gödel datant de 1930, que si ZF est cohérente, alors $ZF \not\vdash \neg AC$. En 1969, Cohen est parvenu à créer un modèle de ZF dans lequel la négation de l'axiome du choix est valide, montrant que si ZF est cohérente, alors $ZF \not\vdash AC$. Ces deux résultats indépendants permettent de conclure que si ZF est cohérente, alors l'axiome du choix est indépendant des axiomes de ZF. Il est possible d'obtenir un résultat analogue avec l'hypothèse du continu : en fait, cette hypothèse est valide dans le modèle de Gödel, mais non dans celui de Cohen. Pour plus de détails, consulter VAN DER LINDEN, *Notions de logique mathématique*, p. 50.

11. Cette illustration de l'axiome du choix n'a pas la prétention de retracer les enjeux véritables que pose la théorie ZFC aux yeux des mathématiciens. La théorie ZFC contient en effet bien plus de résultats contre-intuitifs que la décomposition de la sphère dont nous avons parlé. En fait, les termes mêmes de la théorie ZF sont abstraits : ils traitent

L'axiome du choix est cependant indispensable à la démonstration d'un grand nombre de théorèmes, à commencer par le théorème dit de Zermelo, stipulant de tout ensemble qu'il peut être bien ordonné. Ce résultat est fondamental en théorie des ensembles et représentait d'ailleurs l'objectif visé par Cantor au moment d'étudier les premières classifications des ensembles. Il fournit en effet la garantie (1) qu'il existe un ensemble minimal (l'ensemble vide) et (2) que tous les ensembles peuvent être classés en fonction de leurs tailles. Pour tout ensemble X et Y de taille finie ou infinie, il est possible de dire si $X < Y$, si $X = Y$ ou si $X > Y$. L'on dit alors des symboles $>$, $<$, $=$ qu'ils définissent une relation d'ordre.

L'axiome du choix est aussi nécessaire pour la preuve de différents résultats d'algèbre et d'analyse, parmi lesquels^[12] :

- **Le théorème de Krull** : tout anneau unitaire a des idéaux maximaux ;
- **Le théorème de Tychonoff** : un produit d'espaces compacts est compact ;
- **Le théorème de Nielsen–Schreier** : tout sous-groupe d'un groupe abélien libre est abélien libre ;
- **L'existence d'ensembles non mesurables** (au sens de Lebesgue) ;
- **Le théorème de Hahn–Banach** : l'existence d'une forme linéaire sur un espace de Banach vérifiant certaines conditions ; les valeurs sont imposées sur une partie de l'espace, mais limitées partout ;
- **Le théorème de Stone** : dualité entre la catégorie des algèbres de Boole et celle des espaces topologiques de Hausdorff compacts et totalement discontinus.

Il permet enfin d'établir quelques propriétés élémentaires dans $\mathbb{R}$, comme l'équivalence entre deux définitions usuelles de fonctions continues : celle en termes d'épsilon-delta et celle en termes de suites convergentes^[13]. Compte tenu de tout ceci, peut-on, en toute conscience morale, renoncer à l'axiome du choix ? Ou le prix à payer paraît-il au contraire inadmissible ? Dans ce cas, comment fonder les propriétés mathématiques recensées plus haut ?

d'ensembles d'objets tout à fait *quelconques*, qui ne renvoient pas directement à une réalité intuitive palpable. C'est pourquoi lorsqu'ils en font usage, les mathématiciens ne se soucient pas tant de leurs interprétations intuitives que de leur utilité dans le cadre des mathématiques. C'est une question propre aux philosophes que de demander la signification de ZFC. Pour mener cette question de façon rigoureuse, il serait crucial de dépasser l'exemple académique de l'axiome du choix et d'étudier consciencieusement (1) les principaux objets mathématiques figurant dans la théorie, (2) les propriétés de ces objets et (3) les liens d'implication existant entre cette théorie et les autres théories mathématiques.

12. Tiré de VAN DER LINDEN, *Notions de logique mathématique*, p. 48.

13. *Ibidem*. La première définition implique la seconde, mais la réciproque n'est pas vraie dans ZF sans l'axiome du choix.

5.2.4 Les Théorèmes de Gödel

On pourrait croire que le paradoxe de Russell et de Cantor signent l'arrêt de mort du projet d'axiomatiser les mathématiques ? C'est ici qu'intervient un autre logicien. Notre homme est au moins aussi précis, austère, obstiné et intègre que Gottlob Frege, mais avec un petit je-ne-sais-quoi de plus torturé, de plus tourmenté. Kurt Gödel, d'origine autrichienne, est connu pour sa haute sensibilité. Pour sa crainte des choses invisibles : les bactéries, les atomes et tout ce qui peut susciter l'hypocondrie. À Princeton, c'est le meilleur ami d'Einstein. Il passera la fin de sa vie à tenter de prouver l'existence du Diable. C'est aussi une des rares personnes à avoir lu les *Principia Mathematica* de Russell et Whitehead. En entier.

Sa réputation de génie tient à deux démonstrations concernant l'incomplétude de l'arithmétique — entendez l'impossibilité de prouver certaines vérités à propos de l'addition, de la soustraction et de la multiplication des nombres entiers. Ces démonstrations apportent une réponse au problème de Hilbert concernant la question de savoir si on peut prouver toutes les propositions vraies grâce à un processus automatique consistant à dériver des équations sur une feuille de papier. Hélas pour les tenants de ces authentiques machines de papier et de crayon, la réponse est négative. Pour le prouver, Gödel s'inspire du célèbre paradoxe d'Épiménide. Ce paradoxe consiste à envisager la phrase P disant à propos d'elle-même :

La phrase P est fausse.

Si cette phrase est fausse, elle doit être vraie. Mais si elle est vraie, elle doit être fausse. Ce paradoxe est connu depuis l'Antiquité. Les traités de logique anciens sont truffés de menteurs disant, à propos d'eux-mêmes, qu'ils mentent toujours. Ces jeux de langage, en dépit de la curiosité qu'ils ne manquent pas de susciter, n'étaient pas considérés comme des vérités dignes d'attention. La grande idée de Gödel consiste à s'inspirer de cette phrase P paradoxale pour montrer qu'il existe des vérités arithmétiques qui sont vraies, mais rigoureusement indémonstrables.

L'idée de départ de Gödel consiste à construire un système d'axiomes permettant de construire une phrase P disant à propos d'elle-même « P est indémontrable ». Si on parvient à démontrer P , il s'ensuit que P est faux et ne peut pas être démontrée, ce qui est contradictoire. Par suite de quoi il faut que P soit vraie mais indémontrable. Ce raisonnement est d'un intérêt capital : il suggère que le concept de vérité, au sein du système d'axiomes A , n'est pas équivalent au concept de démontrabilité.

Avant les résultats de Gödel, les partisans du formalisme de Hilbert pensaient (ou voulaient croire) que toute proposition vraie était démontrable, et que toute proposition fausse était telle que sa négation puisse être démontrée. Or Gödel vient établir qu'il existe des propositions vraies indémonstrables, amorçant un divorce complexe et nuancé entre la sémantique (le vrai et le faux) et la syntaxe (le démontrable).

Pour que l'idée de Gödel puisse fonctionner, il faut que le système d'axiomes A , permettant de construire la phrase P , respecte au moins deux conditions : tout d'abord, (1) ce système doit exister et (2) ce système doit être intéressant mathématiquement. La condition (1) signifie qu'il faut trouver des axiomes permettant de construire une phrase qui parle de sa propre démontrabilité. Cette tâche est loin d'être triviale. Au sein d'un système d'axiomes tel que l'arithmétique de Peano, les objets que l'on manipule correspondent à des nombres entiers naturels, à des opérations d'addition ou de soustraction, etc. Ces objets ne réfèrent pas directement à la notion de démontrabilité. Cette notion n'est en effet pas une propriété de l'arithmétique elle-même, mais une propriété du système représentant l'arithmétique. On dira, formellement, que la démontrabilité est une méta-propriété.

La condition (2) stipule que notre système d'axiomes A , en plus de parler de sa propre démontrabilité, doit également être intéressant mathématiquement. Ce que je veux dire par là, c'est qu'il doit être assez puissant pour pouvoir parler des objets mathématiques tels que ceux qui sont manipulés en arithmétique. Si ce n'était pas le cas, le résultat de Gödel pourrait être balayé d'un revers de main, et regardé comme une simple « curiosité logique ». Il n'y a en effet pas beaucoup d'intérêt à savoir qu'un certain système d'axiomes que personne n'utilise, et qui ne peut pas servir à exprimer des vérités mathématiques pertinentes, comme le théorème de Fermat ou la conjecture de Goldbach, possède la propriété étrange d'avoir des propositions vraies et indécidables.

Pour résoudre le point (1), Gödel crée un système d'axiomes permettant de représenter toutes les formules logiques, du type « $0 = 0$ », par des nombres entiers univoques. Si l'on détient deux nombres différents, l'on sait qu'ils désignent deux formules distinctes. Inversement, si l'on détient deux formules logiques distinctes, l'on sait qu'elles correspondent à des nombres différents. Cette approche est connue, en cryptographie, sous le nom d'encodage.

Si l'encodage est suffisamment bien réalisé, il devient alors possible de réduire la démontrabilité d'une phrase à un calcul arithmétique. Une formule du type « $0 = 0$ » sera démontrée vraie s'il existe un calcul sur les nombres permettant d'arriver à la conclusion. De la même manière, une formule « $0 = 1$ » sera démontrée fausse s'il existe un calcul qui montre son contraire, à savoir « $0 \neq 1$ ».

Dans ce paradigme, il devient possible d'exprimer la phrase P disant « La phrase P n'est pas démontrable. » Cette phrase P est un nombre de l'arithmétique et la question de savoir si P est démontrable consiste en un simple calcul. Si l'on parvient à produire ce calcul, cela signifie que P est démontrable. Or comme elle dit d'elle-même qu'elle n'est pas démontrable, cela entraîne une contradiction. Il s'ensuit que P est effectivement indémontrable, mais qu'il n'est pas possible de le prouver.

Notons que ce système d'axiomes, produit par Gödel, est capable de

faire des calculs sur les nombres entiers. Il est donc assez puissant pour exprimer diverses propositions arithmétiques, comme par exemple le théorème de Fermat ou la conjecture de Goldbach. La condition (2) est donc bel et bien satisfaite.

Cette stratégie adoptée par Gödel lui permet de prouver le premier théorème d'incomplétude qui porte son nom :

Théorème 3 (Théorèmes d'incomplétude de Gödel). *Soit un système d'axiomes A suffisamment fort pour exprimer les propositions de l'arithmétique (ce système peut être celui de Peano ou un système d'une puissance au moins équivalente). A possède au moins une proposition indémontrable (et dont la négation est elle aussi indémontrable).*

Lorsque Gödel a présenté ce théorème pour la première fois, très peu d'auditeurs ont été en mesure de comprendre sa portée philosophique, à l'exception de mathématiciens de premier ordre tels que John von Neumann. Ce dernier lui enverra, dès le lendemain, un billet saluant ses travaux et démontrant un important corollaire. Ce corollaire était déjà connu de Gödel. Il s'énonce de la manière suivante :

Théorème 4 (Théorèmes d'incomplétude de Gödel). *Soit un système d'axiomes A suffisamment fort pour exprimer les propositions de l'arithmétique. La proposition disant « A est non-contradictoire » est une proposition vraie, mais indécidable.*

Loin d'être une simple curiosité logique, les théorèmes de Gödel signent l'échec du projet de Hilbert d'axiomatiser l'arithmétique. On observe en effet un divorce entre la vérité et la démontrabilité. Tout ce qui est vrai n'est pas forcément démontrable du point de vue des axiomes. Tout se passe comme si, après le long détour de la science moderne, l'on était reconduit devant cette espèce de jugement, identifié par Platon, qui sont vrais, mais sans justification.

Les théorèmes d'incomplétude semblent un bon point de départ pour reconstruire la définition du jugement synthétique *a priori* de Kant. Ils plaident en effet en faveur d'une notion de vérité qui n'est pas analytique, au sens où on ne peut pas la démontrer à partir des axiomes. Cette vérité concerne néanmoins des objets de la pensée purs. Elle s'applique aux nombres de l'arithmétique, dont l'existence ne semble pas relever de l'épreuve empirique du monde.

Tout se passe comme s'il existait deux types de concepts : (1) les jugements abstraits définis de façon rationnelle à l'aide de symboles mathématiques et d'axiomes et (2) les jugements intuitifs qui sont vrais sans être démontrables rationnellement. Les premiers jugements portent sur les concepts rationnels et les seconds portent sur une espèce de concepts que l'on pourrait qualifier d'intuitifs. À titre d'exemple, le concept intuitif du nombre entier

est ce à quoi vous pensez lorsque vous énumérez des objets, que vous les comptez, que vous en comparez la quantité. Le concept rationnel du nombre entier est défini par un ensemble d'axiomes tels que ci-dessous :

1. L'élément appelé zéro et noté 0 est un entier naturel.
2. Tout entier naturel n a un unique successeur, noté $s(n)$, qui est un entier naturel.
3. Aucun entier naturel n'a 0 pour successeur.
4. Deux entiers naturels ayant le même successeur sont égaux.
5. Si un ensemble d'entiers naturels contient 0 et contient le successeur de chacun de ses éléments, alors cet ensemble est $\mathbb{N}$.

Idéalement, on aimerait que la définition du concept rationnel coïncide exactement avec notre concept intuitif de nombre. Dans le cas de la géométrie euclidienne, c'est effectivement le cas. Toutes les propriétés que l'on observe en dessinant des figures au crayon et au compas peuvent être prouvées à l'aide d'un ensemble d'axiomes judicieusement choisis. Cette correspondance parfaite entre l'intuition et la raison, entre les modèles manipulables et les axiomes, entre l'intuition et la logique formelle, est ce qu'on appelle la complétude de la géométrie. Cette complétude, du reste, va dans les deux sens : toute vérité intuitive à propos des figures est prouvable et tout ce qui est prouvable est vrai intuitivement.

Ce qui est vrai en géométrie ne semble toutefois pas l'être en arithmétique. La plus grande erreur de la philosophie de Platon a peut-être consisté à voir, dans le cas particulier des cercles et des triangles, le modèle absolu de la rationalité. La quête socratique, qui cherche, qui traque, qui débusque derrière chaque idée une essence rationnelle, s'apparente à la quête du Graal. Après deux mille ans d'atermoiement, les théorèmes de Gödel ont reconduit le paradis extasié des idées pures sur la terre ferme de la désillusion. Il n'est plus permis de dire, dorénavant, que les concepts intuitifs et les concepts rationnels sont les deux faces, les deux visages moqueurs, d'un même sesterce. Au contraire, il faut admettre qu'il y a des concepts intuitifs qui ne sont pas rationalisables et qu'il y a des idées rationnelles qui ne correspondent à aucune intuition humaine. C'est en m'appuyant sur cette idée d'une complémentarité inhérente à la représentation et à la définition, que je reconstruirai le concept de jugement synthétique *a priori* dans le chapitre 8.1 consacré aux concepts intuitifs.

5.2.5 La Vérité est une passoire

Il est légitime de se demander avant de poser des conclusions hâtives, si cette interprétation que j'ai faite des théorèmes de Gödel ne serait pas trop charitable. On pourrait penser que les propositions vraies, mais indémonstrables, ne sont finalement que des cas très marginaux au sein de

l'arithmétique. La phrase P de Gödel, en effet, n'est qu'une phrase sans contenu, qui parle d'elle-même dans le vide de l'espace logique. Elle ne dit rien de substantiel à propos des mathématiques. Cette phrase, du reste, est peut-être éliminable ? On peut se demander si, en ajoutant des axiomes, il ne serait pas possible de contourner l'incomplétude identifiée par Gödel ? Je résume ces objections ci-dessous, sous la forme de trois grandes questions :

1. Est-il possible de compléter l'arithmétique en ajoutant la proposition indécidable de Gödel aux axiomes de la théorie ?
2. Les propositions indécidables identifiées par les théorèmes de Gödel ont-elles une signification arithmétique ou ne s'agit-il que de phrases vides de contenu ?
3. Existe-t-il d'autres propositions indécidables que celles relevées par Gödel ?

L'objection (1) consiste à penser que, dès lors qu'une proposition indécidable est rencontrée, il suffit de stipuler dans les axiomes si elle doit être tenue pour vraie ou pour fausse. Dès lors qu'elle figure dans les axiomes, en effet, la proposition est trivialement démontrable. Nous avons rencontré cette approche au moment d'analyser les *Éléments* d'Euclide. On a vu que le cinquième postulat est en effet indécidable eu égard aux quatre autres postulats ; autrement dit, il n'existe aucun moyen de le prouver ou de l'infirmer. Il en découle que l'on peut soit tenir ce postulat pour vrai, soit le tenir pour faux. Dans le premier cas, on obtient la géométrie euclidienne. Dans l'autre cas de figure, cela produit, comme on l'a vu, des géométries non euclidiennes.

Cet artifice ne permet pas, hélas, de venir à bout du problème. Si on ajoute la proposition P de Gödel aux axiomes A de l'arithmétique, on obtient un nouveau système A' d'axiomes. Ce système A' répond toujours aux prémisses du premier théorème de Gödel : il permet en effet de formaliser l'arithmétique. Il est donc possible de reconstruire, au sein du système A' , une nouvelle proposition de Gödel P' qui est vraie, mais indécidable. L'incomplétude revient en force, comme la marée qui se lance à l'assaut de nos modestes châteaux de sable. Mais peut-être dira-t-on qu'il ne faut pas baisser les bras devant le vent et la poussière ? Qu'il faut répéter le raisonnement jusqu'à ce que la mort s'ensuive ? C'est hélas peine perdue. Car si on ajoute à nouveau cette proposition P' aux axiomes de la théorie, on obtiendra un système d'axiomes A'' de laquelle sortira une proposition de Gödel P'' , et ainsi de suite. Il n'y a donc aucune stratégie permettant, à l'issue d'une succession de choix, d'obtenir un système complet. Pour le dire autrement, les axiomes de l'arithmétique contiennent un trou qu'il est impossible de colmater, sauf à accepter un nombre infini d'axiomes.

Si l'on est forcé d'admettre la présence de la proposition indémontrable de Gödel, il reste toutefois possible d'en minimiser la pertinence. L'objection (2) consiste à dire que la proposition P de Gödel, aussi irrésistible soit elle,

n'a pas de signification pour l'arithmétique. Une proposition disant d'elle-même « Je suis indémontrable » apparente davantage à un artifice de langage qu'à une vérité arithmétique. Cette objection que j'esquisse ici, en quelques phrases succinctes, était en vigueur dans le milieu scientifique bien avant que ne soient connus les théorèmes de Gödel. Le mathématicien Henri Poincaré, pour ne citer que lui, disait à propos du paradoxe de Russell : « La logique n'est plus stérile : elle engendre des contradictions¹⁴ ! »

Pour répondre à cette affirmation de Poincaré, j'aimerais faire remarquer deux choses. La première remarque, c'est que le second théorème de Gödel, loin d'être stérile, semble dire quelque chose à propos de l'arithmétique. Il dit en substance qu'il est impossible de prouver qu'elle ne se contredit pas. Ce résultat, si on veut bien le regarder dans le blanc des yeux, est proprement extraordinaire. Depuis la nuit des temps, en effet, l'idée que « $2 + 2 = 4$ » semblait si bien couler de source que l'expression est devenue un synonyme d'évidence logique, un idéal de vérité éternelle, un parangon de raisonnement logico-déductif. J'entends toutefois bruisser la langue de mes détracteurs dans les coulisses. Peut-être dira-t-on que le deuxième théorème de Gödel n'est pas d'une vérité arithmétique *en soi*, mais une vérité *à propos* de l'arithmétique. Une méta-propriété.

Pour fermer le claquet à toute controverse, et c'est là la deuxième remarque que je voulais faire, il convient de souligner que, depuis la publication des théorèmes de Gödel, des propositions arithmétiques indécidables ont été découvertes. Youri Matiassevitch, s'appuyant sur les travaux préalables de Julia Robinson et de Martin Davis, est parvenu à montrer, en 1970, qu'il n'existe pas d'algorithme général permettant de trouver les racines entières d'un polynôme. Qu'est-ce que j'entends par là ? On sait que certaines équations du type $x + 1 = 0$ admettent des solutions entières (sans virgules). On sait également que des équations du type $x^2 + 2y^2 = 0$ n'admettent pas de solution entière. Au début du vingtième siècle, le mathématicien Hilbert a soulevé la question de savoir si l'on pouvait trouver une méthode de calcul permettant de dire, sans erreur possible, si une équation polynomiale à n inconnues et à coefficients entiers possédait des solutions entières ? Ces équations polynomiales à coefficients entiers sont de la forme suivante :

$$a_1x + a_2x^2 + \dots = 0$$

où les coefficients $a_1, a_2, \dots$ sont des entiers (sans virgules). On exclut donc les équations du type :

$$\sqrt{x} + \frac{y}{z} = 0$$

Ce problème, connu sous le nom du dixième problème de Hilbert, a reçu une réponse négative après les publications de Matiassevitch. Il s'ensuit que,

14. HENRI POINCARÉ, *Les mathématiques et la logique*, p.306.

quelle que soit la méthode de résolution que vous envisagez, il est possible d'isoler au moins une équation pour laquelle votre méthode échouera à apporter une réponse satisfaisante en un temps fini. Ce résultat d'indécidabilité conduit à une réécriture très intéressante du théorème de Gödel.

Théorème 5. *Chaque fois qu'on construit un ensemble d'axiomes permettant de formaliser l'arithmétique, il est possible de trouver une équation polynomiale à coefficients entiers qui n'admet aucune solution entière. Ce fait est vrai, mais il ne peut pas être démontré dans l'axiomatisation en question.*

Notons que les résultats de Matiiasevitch ne sont pas isolés. Depuis les années 40, de nombreux problèmes indécidables ont été découverts dans le champ des mathématiques. On peut considérer, à titre d'exemple, le *problème des matrices mortelles*, qui consiste à déterminer si, en multipliant et en additionnant un ensemble de matrices de dimensions $n \times n$ à coefficients entiers, il est possible d'obtenir la matrice nulle composée que de zéros. Un autre exemple est le *problème de correspondance de Post*. Dans ce dernier, on considère deux cellules contenant chacune n brins d'ADN distincts. On se demande si les cellules peuvent, chacune, juxtaposer ces brins d'ADN de sorte à produire la même séquence d'ADN finale. Il s'agit donc de construire un même objet avec des parties différentes.

Ce nombre croissant de problèmes indécidables nous reconduit face à l'objection (3). Celle-ci, rappelons-le, consiste à se demander si les propositions indémontrables sont nombreuses ou s'il ne s'agit jamais que d'un phénomène proprement marginal ? On pourrait penser en effet que l'indécidabilité n'est jamais qu'un comportement anecdotique, affectant les frontières du savoir mathématique, sans menacer le projet général de rationaliser les vérités intuitives au moyen de l'axiomatisation.

Il se trouve qu'il n'en est rien. Au cours de ces dernières décennies, d'importants progrès ont été réalisés pour quantifier d'une manière ou d'une autre l'information que peut nous apporter un système formel donné. Ceci a conduit à une interprétation plus vaste et plus profonde des théorèmes d'incertitude. Une première tentative a notamment été entreprise par Chaitin^[15]. Il énonce un principe heuristique selon lequel les théorèmes prouvés par un système formel donné ne peuvent pas être significativement plus "complexes" que la théorie elle-même.

Calude et Jürgensten ont montré que pour tout système formel S , il existe une constante κ telle que si la complexité d'une formule logique excède κ , alors la formule est indémontrable^[16]. La façon dont il faut entendre la notion de « complexité » fait l'objet d'une définition rigoureuse. Elle est liée au nombre de symboles minimums qu'il est nécessaire d'utiliser pour exprimer la formule dans un langage rugueux.

15. Voir CALUDE et JÜRGENSTEN [?], *Is complexity a source of Incompleteness ?*, p. 1.

16. *Ibidem*, p. 9.

Une question intéressante, à ce stade, est celle de savoir quelle est la proportion de propositions qu'une théorie peut démontrer ? Pour y répondre, les auteurs commencent d'abord par montrer qu'il existe une infinité de propositions ϕ telles que $\delta(\phi) > \kappa$, ou en d'autres termes, qu'il existe une infinité de propositions qui ne sont pas des théorèmes de la théorie. En poursuivant le raisonnement, ils arrivent à la conclusion qu'au plus les propositions sont longues, au plus elles ont de chances de satisfaire cette condition. Dans le cas limite où leur taille n tend vers l'infini, quasi aucune n'est un théorème. Cela les conduit à une forme généralisée du théorème de Gödel, que j'appelle personnellement *Le Théorème insulaire*. Il consiste en effet à dire que, parmi l'ensemble des vérités mathématiques envisageables, seul un faible îlot, balayé par les vents, est démontrable. Il en va des formules démontrables comme des nombres entiers. Il existe une infinité de nombres entiers, mais au plus on s'approche de l'infini, au plus ils se font rares. Arrivé à un certain point, ce ne sont plus que des particules isolées dans le froid glacé, le vide béant des distances immenses.

Théorème 6 (Calude et Jürgensten, 2005). *Soit un système formel respectant les hypothèses de Gödel. La probabilité qu'une proposition ϕ vraie de taille n soit dérivable dans la théorie tend vers zéro lorsque n tend vers l'infini.*

S'il est permis de résumer l'état actuel du projet logiciste en quelques mots, je dirais que la situation contemporaine illustre assez bien la tension entre les propositions analytiques et l'intuition. Bien loin de capturer précisément l'intuition des théories mathématiques, le développement d'une axiomatique rigoureuse semble révéler chaque jour un peu plus l'autonomie relative de la pensée logique par rapport à la pensée intuitive. Tout se passe comme si l'être humain était capable de penser des totalités que la logique ne permettait pas d'exprimer.

Le plus grand paradoxe logique, cela étant dit, vient peut-être de ce que, en dépit de la grande impuissance des procédures de calcul et du nombre finalement très limité de vérités rationnellement appréhendables, on ne continue pas moins d'utiliser la raison calculante et l'ordinateur, comme s'il s'agissait des seules voies d'accès légitimes au savoir.

5.3 Les Théories de la calculabilité

L'on en vient incidemment au deuxième paradigme de l'informatique balbutiante, qui consiste à axiomatiser le raisonnement rationnel en construisant un modèle de calcul automatique. Ce modèle doit fonctionner de telle sorte que, pour toute question posée, et pour peu que cette question soit suffisamment précise pour recevoir une réponse univoque, on obtienne une réponse valide. La raison peut, dans cette perspective, être modélisée à la

façon d'une boîte noire, associant un input à un output de façon déterministe.

Pour définir en des termes précis le concept de calculabilité, Hilbert se référait à la fiction d'un homme dérivant des symboles sur un bloc de feuilles. Si l'on prend au sérieux cette conception, mais en faisant l'abstraction des tendons, des nerfs et de toute cette ossature biologique, somme toute contingente, constitutive de l'être humain, nous pouvons réduire cette idée à son plus simple aspect : d'un côté, nous avons un ruban de papier, décomposé en un certain nombre de cases ; de l'autre côté, une tête de lecture mécanique capable de circuler librement sur le ruban, de s'arrêter sur un emplacement, de lire le symbole qui s'y trouve (ou de constater l'absence de symbole), de l'effacer et d'en écrire un nouveau.

Tout calcul consisterait dès lors, pour la machine ainsi obtenue, à lire une question qu'on lui fournit sur le ruban au moyen d'un certain nombre de symboles et à offrir une réponse sur ce même ruban avec des symboles similaires. Le type de symboles admissibles est décrit par un alphabet Σ . Dans sa forme la plus simple, l'alphabet est binaire et n'est donc constitué que de 0 et de 1. Dans sa forme la plus sophistiquée, l'alphabet peut contenir un nombre infini de symboles. Il suffit de penser, à titre d'exemple, à l'ensemble $\mathbb{N}$ des naturels (infini dénombrable) ou à l'ensemble $\mathbb{R}$ des réels (infini indénombrable). Quel que soit la nature de l'alphabet, on utilisera le symbole $\sqcup$ pour représenter le fait qu'une case donnée du ruban est vierge. Le symbole $\sqcup$ est donc un symbole réservé, ou encore un métasympbole, qui ne peut pas figurer parmi les autres symboles de l'alphabet Σ .

Si la taille du ruban est « suffisamment grande », alors tout alphabet peut être représenté par la machine de façon binaire. Concrètement, la taille de l'alphabet ne change dès lors pas le type de questions que la machine peut lire ni le type de réponses qu'elle est capable de fournir. Par l'expression « suffisamment grand », j'entends un ruban dont la taille peut être prolongée à l'envi. Un tel ruban correspondrait à la situation imaginée par Hilbert lorsqu'il dit du calculateur qu'il a le droit d'utiliser autant de feuilles de papier qu'il le souhaite. À tout moment, la quantité de feuilles utilisées peut être décrite avec un nombre fini, mais ce nombre lui-même ne connaît pas de borne.

Une autre propriété importante, nonobstant la taille du ruban, est la mémoire de la machine. Dans l'hypothèse de Hilbert, le mathématicien ne se contente pas de lire les symboles sur une feuille de papier, mais il peut aussi annoter ces feuilles au cours de sa lecture ; voire même noter ses réflexions sur un calepin à part. Une machine qui, à l'instar de ce calculateur humain, possède la propriété de garder des éléments en mémoire, pourra répondre à des questions plus complexes que celle qui se contente de parcourir le document symbole par symbole, sans rien mémoriser.

Les machines possédant une mémoire finie sont appelées les *automates cellulaires* et il en existe de plusieurs espèces : celles dotées d'un calepin

externe et celles qui se contentent de lire un ruban de taille finie en répondant par "oui" ou par "non" au terme du processus. Les machines possédant une mémoire indéfinie, quant à elles, qui ont en outre la possibilité d'écrire des réponses plus élaborées qu'un simple "oui" ou "non", sont les *machines de Turing*.

Bien que l'usage de l'ordinateur nous ait habitué à l'idée qu'une machine possède une capacité de mémoire limitée, je vais montrer à l'issue de cette section que le modèle le plus adéquat pour représenter la notion de calculabilité est la machine de Turing. L'argument se décompose en un versant minimaliste et en un versant maximaliste.

Le versant minimaliste consiste à exhiber une fonction qui, bien qu'incalculable par un automate cellulaire, l'est par une machine de Turing (et le serait par un humain avec une feuille de papier aussi longue que souhaitée). Il faut donc au moins un modèle aussi puissant que la machine de Turing pour capturer le concept de calculabilité.

Le versant maximaliste, quant à lui, consiste à dire qu'il n'existe aucun modèle fini plus puissant que la machine de Turing. Autrement dit, toute modification à laquelle on pourrait songer pour améliorer la machine de Turing, en la dotant par exemple de bandes de mémoire supplémentaires ou de têtes de lecture plus sophistiquées, augmenteraient certes sa rapidité de traitement, mais pas son pouvoir de calcul réel. La machine de Turing est ainsi le plus petit modèle capturant entièrement toutes les capacités de calcul d'un ordinateur.

Il est important de noter ici la distinction existant entre une machine réelle, comme un ordinateur, et la machine de Turing. L'ordinateur n'est pas un modèle mathématique, mais un objet physique doté, en plus d'une certaine organisation structurelle, de câbles et de signaux électriques (hardware). S'il fallait modéliser l'ordinateur, il faudrait recourir à une machine de Turing avec un ruban de taille finie. Cette dernière occuperait une place intermédiaire entre l'automate cellulaire, qui a aussi un ruban fini mais ne peut pas l'éditer, et la machine de Turing proprement dite.

La machine de Turing avec un ruban infini, quant à elle, n'est pas un modèle de l'ordinateur, mais de la notion abstraite de calculabilité. Elle est un modèle de la calculabilité, au même titre que le triangle mathématique est un modèle de la géométrie. Il n'existe sans doute pas, comme le faisait remarquer Descartes, de solides parfaitement triangulaires dans le monde, pas plus qu'il n'existe d'ordinateur à mémoire parfaitement infinie. Néanmoins, étudier les propriétés des triangles nous enseigne quelque chose sur les solides de formes grossièrement triangulaires qui se rencontrent de par le monde. De la même manière, étudier les propriétés d'une machine de Turing nous renseigne sur les forces et les limitations des processus de calculs — et donc indirectement sur certaines propriétés de l'esprit humain (supposition faite que l'esprit humain consiste, au moins en partie, en quelque chose de calculable).

5.3.1 Formalisme de la Machine de Turing

Alphabet

Avant de poursuivre plus avant, il convient de poser quelques définitions formelles. Un ensemble de symboles Σ est appelé un alphabet, qu'il soit de taille finie ou non. Les alphabets les plus courants sont l'alphabet binaire $\{0, 1\}$ et l'alphabet latin $\{a, b, c, \dots, y, z\}$. Il n'y a toutefois pas de restriction sur les symboles que peut contenir Σ . Souvent, le caractère $\sqcup$ est utilisé pour représenter un espace blanc.

Une *chaîne de caractères*, aussi appelée un *mot*, est une suite de symboles de cet alphabet. On peut la représenter au moyen d'une séquence $\sigma_1, \sigma_2, \sigma_3, \dots$, où σ_1, σ_2 et σ_3 appartiennent à Σ . On dira donc que 100101 est une chaîne de caractères de $\{0, 1\}$ et que *logique* est un mot de l'alphabet latin. Sous cette convention, le symbole 0 représente à la fois un élément singulier de l'alphabet et la chaîne constituée de cet unique élément. Par ailleurs, on notera "" la chaîne de caractères vide.

La longueur d'une chaîne de caractères s est le nombre de symboles dont elle est constituée. Elle sera souvent notée $|s|$. Par exemple, $|machine| = 7$. Au besoin, on notera $s(i)$ le symbole situé au i -ème emplacement de la chaîne s , si bien que, dans l'exemple ci-dessus, $s(1) = m$ et $s(6) = n$.

Deux chaînes de caractères x et y définies sur un même alphabet peuvent être concaténées au moyen de l'opérateur ".". Le produit de cette opération est le mot obtenu en ajoutant les symboles de y à la droite de x tout en conservant l'ordre à l'identique. En d'autres termes, $100101 = 10 \cdot 0101$ et *parapluie* = *para* · *pluie*. Lorsque le contexte est suffisamment clair, on peut omettre le symbole de concaténation et écrire simplement xy au lieu de $x \cdot y$. Dans tous les cas de figure, x est appelé le préfixe et y est le suffixe.

L'inverse d'une chaîne de caractères x est notée x^R . Elle consiste à écrire chacun de ses symboles dans le sens inverse. L'inverse de *repas* est donc *saper*. Dans le cas particulier où $x = x^R$, on parle de palindrome, comme dans *ressasser*, *radar* ou 1001.

Langage

Étant donné un alphabet Σ , on note Σ^* l'ensemble des mots qu'il est possible de former dans cet alphabet par concaténation de ses symboles (on appelle * l'opération de Kleene). Dans le cas de l'alphabet latin, Σ^* contient tous les mots possibles ou imaginaires qu'il est permis de construire. La plupart d'entre eux sont imprononçables, et d'une longueur que n'aurait rien à envier le serpent mythologique de Jörmungandr, mais une petite minorité de ces mots appartiennent à la langue française. Au vu de ce que nous venons de dire, un *langage* est donc un sous-ensemble des chaînes de caractères, ou mots, qu'il est possible de former dans un alphabet donné.

Si la langue française comprend un nombre fini de mots, susceptibles d'être indexés dans un dictionnaire, ce n'est pas le cas de la plupart des langages étudiés en théorie de la calculabilité. Par exemple, le langage L des palindromes est l'ensemble infini des mots pouvant se lire aussi bien de gauche à droite que de droite à gauche. On écrira $L = \{x \in \Sigma^* \mid x = y \cdot y^R, \text{ avec } y \in \Sigma^*\}$. Dans cette expression, la propriété $x = y \cdot y^R$ est ce qui distingue les palindromes des autres mots de Σ^* .

Notons au passage que le fait même de construire un langage L à partir d'une propriété, sans spécifier la procédure permettant d'énumérer méthodiquement les mots de ce langage, est ce que Poincaré appelait l'infini extensif. Infini, car on obtient souvent des langages de taille infinie ; extensif, car le langage ainsi obtenu n'est, ni plus ni moins, l'extension des mots qu'il comporte.

Il est facile, en s'appuyant sur l'infini extensif, de produire des paradoxes. Pour ce faire, supposons, à titre préalable, que tous les mots formés à partir de l'alphabet latin Σ , auquel on ajoute l'espace blanc $\sqcup$, soient triés par ordre lexicographique croissant. On aurait dans ce classement d'abord les mots de longueur un commençant par a , puis par b , par $c...$ après quoi on aurait les mots de longueur deux commençant par a , puis par b , et ainsi de suite. Définissons ensuite un langage L formé de tous les mots ayant la propriété suivante : « Le mot est situé avant le premier mot de la liste lexicographique croissante qui ne soit pas exprimable en moins de 132 caractères. »

On remarque que l'expression entre guillemets contient moins de 132 caractères, espaces compris. Dès lors, soit tous les mots de notre ordre lexicographique croissant sont exprimables en moins de 132 caractères, dans quel cas L contient tous les mots possibles et ceux-ci sont en nombre fini ; mais cela contredit l'idée que les mots possibles soient des concaténations infinies dénombrables des lettres de l'alphabet. Soit il faut qu'en énumérant les mots de notre classement lexicographique croissant, on finisse par trouver l'un d'eux qui ne soit pas exprimable à l'aide de 132 caractères ; mais ce mot l'est pourtant grâce à la phrase entre guillemets. Le langage L est donc bien paradoxal^[17].

Ce qui rend ce langage paradoxal, c'est que la propriété P sur laquelle il repose fait référence à un certain ordonnancement des mots. Or sitôt que l'ordonnancement est construit, il est possible d'y faire référence pour obtenir un nouveau mot avec moins de mots qu'il en était nécessaire dans le premier référencement. Cela aboutit à un nouveau référencement, et ce indéfiniment.

Pour éviter l'irruption de cette espèce de paradoxes, il peut être préférable de n'envisager que les langages pouvant être construits dans l'alpha-

17. Ce paradoxe est une reformulation du paradoxe de Berry, introduit sous ce nom par Bertrand Russell en l'honneur du bibliothécaire londonien éponyme. Le paradoxe consiste à envisager : « plus petit entier non nommable en moins de dix-huit syllabes qui paraît être ainsi nommé en dix-sept syllabes. »

bet uniquement au moyen de l'opération de concaténation, de l'opération de Kleene $*$ et en autorisant l'union de plusieurs langages obtenus de cette manière. On dira de tels langages qu'ils sont réguliers et, par extension, que leurs mots sont des expressions régulières. Un exemple de langage régulier est le langage $\mathcal{L} = \{w \in \{0,1\}^* : w \text{ a deux ou trois occurrences de } 1, \text{ les deux premières n'étant pas consécutives}\}$. Ce langage est régulier car il est possible de donner la règle de construction des mots qu'il comprend en utilisant uniquement les règles usuelles de concaténation et l'opération $*$. Il s'agit des mots obtenus comme ceci : $0^* \cdot 1 \cdot 0 \cdot 0^* \cdot 1 \cdot \{10\}^*$.

Lorsqu'un langage est régulier, il est facile de déterminer si un mot appartient à son langage. Dans le cas contraire, il est permis de se demander s'il existe un procédé algorithmique permettant de déterminer si une suite de symboles est, oui ou non, un mot d'un certain langage. L'existence d'un tel procédé donne la garantie que le langage en question est constructif. Dans le langage de Hilbert, cela signifierait que le langage peut être obtenu par une certaine manipulation de symboles. La manipulation la plus simple que nous puissions imaginer, à ce stade, est celle de l'automate cellulaire.

Automate cellulaire

Un automate cellulaire est une tête de lecture capable de lire le contenu d'un ruban, case par case, et de répondre à la fin de la lecture par "oui" ou par "non". On dit dans le premier cas que le contenu du ruban (l'input) est accepté; dans le cas contraire, on dit qu'il est rejeté. Une telle machine ne peut donc pas revenir en arrière avec sa tête de lecture, ni changer le contenu du ruban. Elle permet cependant, comme on va le voir, de dire d'un mot s'il appartient ou non à un langage régulier.

Pour formaliser l'automate cellulaire, il faut le doter d'un ensemble fini d'états $K = \{q_0, q_1, q_2 \dots q_n\}$ ainsi que d'un ensemble d'instructions (un algorithme) stipulant : "si la machine est dans l'état q et lit un certain symbole σ , elle passe dans l'état q' ". Les instructions sont donc des injections complètes de $(K \times \Sigma)$ vers Q .

Au début de la lecture, il faut préciser dans quel état se trouve la machine. On l'appelle l'état initial. À chaque fois qu'elle lira une case du ruban, la machine sera susceptible de changer d'état, en fonction de ce qui est prescrit par l'algorithme. Lorsqu'elle arrive au bout de la bande, elle se trouvera soit dans un état dit terminal, dans quel cas elle répondra par "oui", soit dans un état non-terminal, dans quel cas elle répondra par "non".

Définition 4. *Un automate cellulaire déterministe est un quintuplet $M = (K, \Sigma, \delta, s, F)$ tel que¹⁸ :*

18. HARRY. R. LEWIS, CHRISTOS H., PAPADIMITRIOU, *Element of the theory of calculation*, p.57.

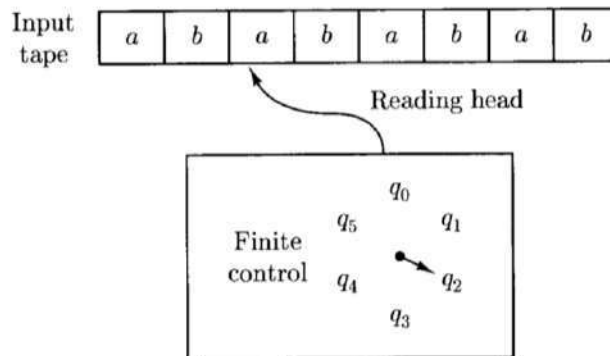


Figure 2-1

FIGURE 5.1 – Schéma représentant un automate cellulaire. Celui-ci est doté d'une bande, d'une tête de lecture et d'un algorithme interne déterminant dans quel état la machine doit se trouver. La lecture de la bande n'est réalisée qu'une seule fois, de gauche à droite. (source de l'image : Lewis & Papadimitriou).

1. K est un ensemble fini d'états $\{q_0, q_1, \dots\}$;
2. Σ est un alphabet de taille finie ;
3. $s \in K$ est l'état initial ;
4. $F \subseteq K$ est l'ensemble des états finaux ;
5. δ la fonction de transition, une injection de $K \times \Sigma \cup \{()\}$ dans K .

Exemple 1. Considérons un automate cellulaire capable de dire d'un mot w s'il appartient ou non au langage $\mathcal{L} = \{w \in \{a, b\}^* \mid w \text{ ne contient pas d'occurrence de } b\}$. Il suffit d'envisager deux états q_0 et q_1 , où q_0 est l'état initial et le seul état terminal. La fonction de transition δ est donnée par la table ci-dessous :

q	σ	$\delta(q, \sigma)$
q_0	a	q_0
q_0	b	q_1
q_1	a	q_1
q_1	b	q_1

Dans cet exemple, l'automate cellulaire $M = (\{q_0, q_1\}, \Sigma, \delta, q_0, q_0)$ commence dans l'état q_0 et lit les cases du ruban une par une. Il passe dans l'état q_1 dès qu'il rencontre un b et n'en sort plus jamais. À la fin de la lecture, il répond "oui" s'il est dans l'état q_0 et "non" s'il est dans l'état q_1 .

Il est possible de représenter la fonction de transition δ d'un automate cellulaire au moyen de la théorie des graphes. Chaque noeud correspondra à un état et les flèches allant d'un noeud à l'autre représentent les transitions d'états autorisés. On prend pour label de ces flèches la condition requise pour

effectuer la transition. C'est ainsi que l'automate de l'exemple ci-dessus peut être représenté à l'aide du diagramme suivant :

Le résultat le plus important concernant les automates cellulaires vient de ce qu'ils sont capables de lire les langages réguliers, et uniquement ceux-là. Il y a donc une identité conceptuelle entre un langage régulier et un automate cellulaire.

Théorème 7. *Un langage est régulier si et seulement s'il est accepté par un automate cellulaire.*

Démonstration. Un schéma de la preuve consiste à montrer que toutes les opérations intrinsèques aux langages réguliers peuvent être acceptées par l'automate cellulaire. On considérera sans perte de généralité un automate cellulaire $M = (K, \Sigma, \delta, q_s, F)$ avec $\Sigma = \{\sigma_1, \sigma_2, \dots\}$.

- **Opération de concaténation :** Lorsque la machine doit vérifier la concaténation de deux lettres $\sigma_1\sigma_2$, elle entre dans l'état q_c . On construit δ de sorte que $\delta(a_2) = q_{oui}$, mais que pour tout autre symbole $\sigma \neq a_2$, on ait $\delta(\sigma) = q_{non}$.
- **Opération de Kleene (*) :** Soit $\Sigma_1 \subseteq \Sigma$. Lorsque la machine doit vérifier que la suite du mot appartient bien à Σ_1^* , elle entre dans l'état q^* . On construit δ de sorte que pour tout $\sigma \in \Sigma_1$, $\delta(\sigma) = q^*$ tandis que pour tout $\sigma' \notin \Sigma_1$, $\delta(\sigma') = q_f$.
- **Union :** Soient $\mathcal{L}_1$ et $\mathcal{L}_2$, deux langages réguliers. On considère, sans perte de généralité, qu'ils sont définis sur un même alphabet Σ (si ce n'est pas le cas, il suffit de redéfinir ces langages sur l'union des alphabets respectifs). Notons $M_1 = (\{q_1, q_2, \dots, q_n\}, \Sigma, \delta_1, q_{s1}, F_1)$ et $M_2 = (\{k_1, k_2, \dots, k_m\}, \Sigma, \delta_2, q_{s2}, F_2)$ les machines acceptant respectivement les mots de ces deux langages. Pour obtenir une machine acceptant $\mathcal{L}_1 \cup \mathcal{L}_2$, il faut construire $M = (K, \Sigma, \delta, q_s, F)$. Pour chaque état q_i et chaque état k_j , on construit l'état conjugué $q_{i,j}$ dans K (avec $i \leq n$ et $j \leq m$). La table de transition de δ est obtenue en conjuguant les tables de transitions de δ_1 et δ_2 . Pour tout état $q_{i,j} \in K$ et pour tout caractère $\sigma \in \Sigma$, la table renvoie donc à l'état $q_{k,l}$, où k est l'indice de l'état $\delta_1(\sigma, q_i)$ et l est l'indice de $\delta_2(\sigma, k_j)$. Un état $q_{i,j}$ est final si et seulement si q_i et k_j sont tous les deux des états finaux.

Ceci prouve qu'il est possible, pour tout langage régulier, de construire une machine acceptant ce langage. Pour prouver la réciproque, considérons automate cellulaire $M = (K, \Sigma, \delta, q_s, F)$ quelconque et montrons que le langage accepté par M est régulier. Considérons sans perte de généralité que $q_1 \in K$ est l'état initial. Notons $R(i, j, n)$ l'ensemble des chaînes de caractères qu'il est possible de lire sur la bande pour que la machine passe de

l'état q_i à l'état q_j sans passer par un état intermédiaire d'indice supérieur à k (il est donc permis de passer par les états $q_1, q_2, \dots, q_k$). Il est facile de voir que les langages acceptés par la machine sont donnés par :

$$\bigcup_{j,n} R(1, j, n) : q_j \in F$$

On sait que l'union de deux langages réguliers est régulière. On va prouver par induction sur n que, pour tout i, j et n , $R(1, j, n)$ est régulier. Pour ce faire, on commence par remarquer que $R(i, j, 0)$ est régulier. C'est assez évident car il s'agit soit d'un langage vide, soit d'un seul caractère. Supposons ensuite que $R(i, j, k)$ soit régulier pour i, j et pour tout $k \leq n$. Les seules chaînes de caractères permettant d'aller de l'état q_i à q_j sans passer par plus des états intermédiaires d'indices supérieures à $k + 1$ vont :

- Soit de l'état q_i à q_j sans passer par un état intermédiaire d'indice supérieur à $k - 1$;
- Soit de l'état q_i à q_k , puis zéro ou plusieurs fois de q_k à q_k , puis de q_k à q_j , le tout sans jamais passer par un état supérieur à k .

On peut donc écrire :

$$R(i, j, k + 1) = R(i, j, k - 1) \cup R(i, k, k).R(k, k, k - 1)^*.R(k, j, k - 1)$$

Comme la concaténation et l'union de Kleene de langages réguliers donnent des langages réguliers, on conclut que $R(i, j, k + 1)$ est régulier pour tout indice i, j et k . $\square$

En guise d'illustration de ce théorème, il est facile de voir qu'il n'existe aucun automate cellulaire capable d'accepter le langage des palindromes $L = \{x \in \Sigma^* \mid x = y \cdot y^R, \text{ avec } y \in \Sigma^*\}$. Il faudrait en effet que la machine mémorise les lettres qu'elle lit au fur et à mesure de façon à vérifier si, une fois arrivé à un certain point de sa lecture, le schéma de lettres mémorisés se répète en sens inverse. Pour ce faire, on pourrait penser à doter l'automate de nombreux d'un état $q_a, q_b, q_c, \dots$ correspondant au cas où la première lettre lue serait un a , un b , un c , ainsi que des états $q_{aa}, q_{ab}, q_{ac}, \dots, q_{ba}, q_{bb}, q_{bc}, \dots, q_{ca}, q_{cb}, q_{cc}, \dots$ permettant à son tour d'enregistrer quelle est la deuxième lettre lue, etc. Mais le problème avec une telle approche vient de ce que le nombre d'états doit être fixé à l'avance. S'il existe n états et que le mot à lire contient trop de lettres, il n'y aura pas assez d'états pour enregistrer tous les cas de figure.

On comprend donc bien que les palindromes de tailles indéterminées ne forment pas un langage acceptable par un automate cellulaire et, par suite, que ce n'est non plus un langage régulier. Cela est cohérent, puisque

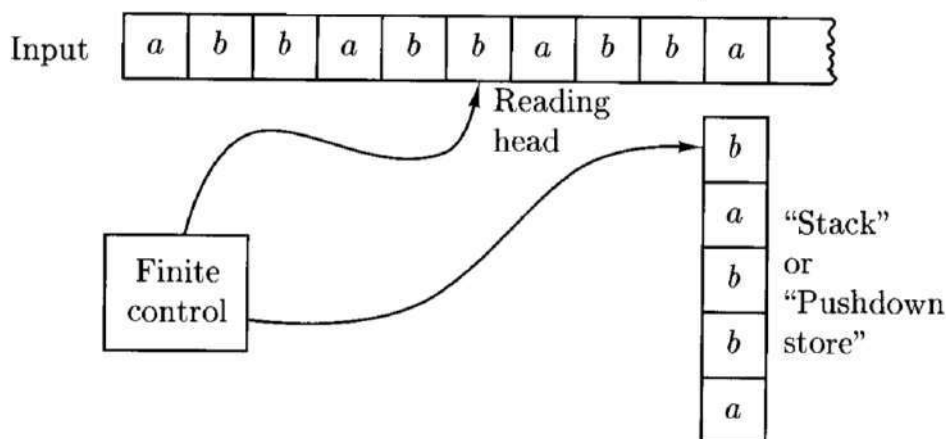


FIGURE 5.2 – Schéma représentant un automate à pile. Cet automate cellulaire amélioré est doté d'une bande externe sur laquelle il peut ajouter ou effacer des symboles. L'ajout d'un symbole se fait toujours par-dessus les symboles existants (opération "pop") et l'effacement consiste à retirer les derniers à avoir été écrits (opération "push"). (source de l'image : Lewis & Papadimitriou.)

l'opération y^R n'est pas une simple opération de concaténation ni une simple opération $*$ de Kleene. Il semble naturel, dès lors, de se demander quelle extension il faudrait ajouter à l'automate cellulaire pour qu'il accepte les palindromes.

Automate à pile

Une simple extension de l'automate cellulaire, capable de lire les palindromes, consisterait à ajouter un ruban annexe à la machine. Ce ruban serait utilisé, à tout moment pendant l'opération, pour noter de l'information et la lire, sans toutefois l'éditer. Le produit de cette adjonction est baptisé sobrement Automate à pile, le mot "pile" était une traduction de l'anglais stack et désigne la capacité à empiler de l'information. Rien à voir donc, avec les batteries de lithium !

Définition 5. *Un automate à pile est un sextuplet $M = (K, \Sigma, \Gamma, \Delta, s, F)$ tel que*

1. K est un ensemble fini d'états $\{q_0, q_1, \dots\}$;
2. Σ est l'alphabet du ruban ;
3. Γ est l'alphabet de la pile ;
4. $s \in K$ est l'état initial ;
5. $F \subseteq K$ est l'ensemble des états finaux ;

6. δ la fonction de transition, une injection de $K \times \Sigma \cup \{''\} \times \Gamma$ dans K^* .

La fonction $\Delta : (q, \sigma, \gamma) \rightarrow (q', \gamma')$ signifie que, si la machine est dans l'état q et qu'elle lit le symbole σ , elle doit passer dans l'état q' . En sus, elle remplace la chaîne de caractères γ du sommet de la pile par la chaîne de caractères γ' . En particulier, la transition $(q, \sigma, '') \rightarrow (q', a)$ signifie qu'il faut ajouter le symbole a au sommet de la pile; on dit que la machine "pop" le symbole a . Inversement, la transition $(q, \sigma, a) \rightarrow (q', '')$ signifie qu'il faut effacer le symbole a du sommet de la pile; on dit que la machine "push" le symbole a . Ce vocabulaire justifie l'appellation alternative de pushdown automata qu'on donne à l'automate à pile.

La Machine de Turing

Outre la tête de lecture et le ruban de taille finie dont nous avons parlé plus haut, une machine de Turing est également dotée d'un *algorithme* définissant, à chaque unité de temps, une opération à exécuter. Les opérations permises consistent à déplacer la tête de lecture d'une case sur le ruban, soit à droite ($\rightarrow$), soit à gauche ($\leftarrow$); à lire le symbole $\sigma \in \Sigma$ figurant sur la case actuelle (on note $\sqcup$ si cette case est vierge) ou à écrire un nouveau symbole $\sigma' \in \Sigma$.

L'algorithme est en quelque sorte le programme de la machine. Il consiste en un certain nombre d'instructions de type *si... alors...* Par exemple : « si le ruban est sur la première case, alors il faut le déplacer sur la deuxième case ». Évidemment, on ne veut pas qu'une machine, lorsqu'elle est sur la première case, aille systématiquement sur la deuxième. On précisera alors "si la tête de lecture est dans l'état q_0 et qu'elle se trouve sur la première case du ruban, alors il faut la déplacer sur la deuxième case; mais si elle est dans l'état q_1 , alors elle doit lire le symbole figurant sur la case, etc." Le nombre d'états $q_0, q_1, q_2, \dots$ autorisés par la machine est illimité, mais borné. Il est donc impossible d'avoir une infinité d'états, tout comme il est interdit d'avoir un algorithme comportant une infinité d'instructions.

En plus de l'état initial, un algorithme prescrit également un état final q_f . Lorsque la machine entre dans cet état, elle s'arrête immédiatement. La seule instruction associée à q_f est donc « si la machine entre dans un état final, quelle que soit la position de sa tête de lecture, elle s'arrête. » Une question importante consiste à se demander si, pour un certain algorithme donné, les questions écrites sur le ruban donnent lieu à une réponse correcte en un temps fini. Pour que cela soit vrai, il faut d'une part que la machine s'arrête; et d'autre part, il faut que la suite de symboles écrite sur le ruban lorsqu'elle s'arrête soit la réponse à la question posée.

Il est facile de voir que certaines machines s'arrêtent pour certaines entrées et ne s'arrêtent pas pour d'autres. Si vous construisez un algorithme

visant à réaliser la division euclidienne, la machine s'arrêtera pour des entrées telles que « divisez 8 par 2 », mais pas « diviser 10 par 3 ». Dans ce dernier cas de figure, la réponse figurant sur le ruban sera 3,33333... avec une suite infinie de nombres consécutifs. Il est toutefois important de noter que si la question est « 10 est-il divisible par 3 ? », il est permis d'écrire un programme qui répond à ce genre de question en temps fini. Ce dernier consiste simplement à parcourir tous les multiples de 3 jusqu'à obtenir 10 (dans quel cas 10 est divisible par 3) ou un nombre strictement supérieur à 10 (dans quel cas 10 n'est pas divisible par 3). Ceci peut s'écrire sous la forme d'un programme :

```

Divisible(s1, s2)
  poser i = s1
  tant que i < s2 :
    si i = s2, répondre vrai ;
    sinon, poser i = i + s1 ;
  sinon répondre faux;

```

Les questions binaires, dont la réponse est sous forme de vrai ou faux, semblent plus faciles à résoudre que les questions ouvertes. En termes techniques, le problème qui consiste à dire si une certaine propriété est vérifiée ou non est appelé *problème de décision*. Un exemple de problème de décision que nous avons déjà vu consiste à se demander si une proposition arithmétique peut être démontrée à partir d'un certain ensemble d'axiomes. On a vu que la réponse était négative dans le formalisme des théorèmes de Gödel. Il est intéressant de se demander si, dans le formalisme des machines de Turing, tous les problèmes de décisions admettent une réponse ?

On dit d'une suite de symboles qu'elle est acceptée par la machine si, lorsque cette suite est écrite sur le ruban, la machine finit par atteindre un état final. On dit au contraire que la suite n'est pas acceptée si la machine ne s'arrête jamais. L'ensemble des suites de symboles x^* acceptées par l'automate cellulaire constitue le langage calculable de cet automate cellulaire. Ces considérations nous conduisent à une définition formelle :

Définition 6. Une machine de Turing est un quintuplet $M = (K, \Sigma, \delta, s, H)$ tel que

1. K est un ensemble fini d'états $\{q_0, q_1, \dots\}$;
2. Σ est l'alphabet du ruban, contenant le symbole blanc $\sqcup$ ainsi que le symbole représentant le début du ruban $\triangleright$ mais pas les symboles $\rightarrow$ et $\leftarrow$;
3. $s \in K$ est l'état initial ;
4. $H \subseteq K$ est l'ensemble des états finaux ;
5. δ la fonction de transition, une injection de $(K - H) \times \Sigma$ dans $K \times (\Sigma \cup \{\leftarrow, \rightarrow\})$.

- (a) Pour tout $q \in (K - H)$, if $\delta(q, \triangleright) = (p, b)$, alors $b = \rightarrow$.
- (b) Pour tout $q \in (K - H)$ et $a \in \Sigma$, if $\delta(q, a) = (p, b)$, alors $b \neq \triangleright$.

5.3.2 Le Problème de l'arrêt

Une importante interrogation consiste à se demander si tous les problèmes sont décidables. Un problème est dit « décidable » si la machine de Turing peut y répondre par vrai ou par faux en utilisant une méthode de calcul univoque. Par exemple, la question de savoir si un nombre est premier est décidable. Il existe en effet un algorithme permettant de le déterminer : celui-ci vérifie, itérativement, si un nombre s donné est divisible par un nombre strictement inférieur à s à part le nombre 1¹⁹. On s'attend toutefois à ce qu'il y ait des problèmes indécidables. Et pour cause : la thèse de Church-Turing stipule que les machines de Turing et les systèmes d'axiomes de Gödel sont deux façons équivalentes d'exprimer le concept de calculabilité. Il s'agirait donc de deux définitions concurrentes d'un même concept intuitif. Cette thèse, bien sûr, ne fait pas l'objet d'une démonstration. Il s'agit seulement d'un consensus, au sein de la communauté mathématique.

Il se trouve qu'Alan Turing est parvenu à déterminer l'existence d'un problème qui est indécidable. Ce problème est connu sous le nom du « problème de l'arrêt ». En des termes intuitifs, il consiste à construire une machine de Turing, appelée *Halt*, qui analyse le fonctionnement de toutes les autres machines de Turing possibles et imaginables de sorte à déterminer dans quelles conditions ces machines s'arrêtent après un nombre fini d'opérations et dans quelles conditions elles bouclent infiniment. Notre machine devrait donc être en mesure de reconnaître qu'un programme qui fait la division euclidienne s'arrêtera lorsqu'il cherche à diviser 10 par 2 mais pas s'il cherche à le diviser par 3.

La création de *Halt* n'est pas du tout évidente. *Halt* doit en effet comprendre ce que font les machines qu'il analyse de façon à déterminer si elles se sont lancées, ou non, dans un processus sans fin. Cette compréhension ne peut pas être obtenue en regardant simplement les symboles que ces machines écrivent sur leurs bandes de lecture. Si on observe un programme qui écrit par exemple les nombres suivants 1421421421... on pourrait penser qu'il ne va jamais s'arrêter de radoter jusqu'à la fin des temps. Il se pourrait toutefois que, après 100 étapes, le programme s'arrête ou change totalement de registre. Pour en avoir le cœur net, un être humain ira consulter le code de la machine. Si ce code a la forme suivante, il saura que le programme s'arrêtera en un temps fini :

```
procedure T(s):
  écrire 142;
```

19. Donald Knuth fournit un algorithme plus élaboré dans son livre intitulé *The Art of Computer Programming : Fundamental Algorithms*, Addison-Wesley, 1997.

répéter cent fois;

Si par contre le code est écrit tel que voici, il saura que le programme continuera sans fin :

procédure $T(s)$:

```
écrire s;
si s est pair, diviser s par deux;
si s est impair, multiplier s par trois et ajouter un;
si s est égal à 100, s'arrêter;
sinon répéter;
```

Dans ce dernier code, en effet, la machine ne s'arrête que si le nombre s , à un moment donné, est égal à 100. Cela peut se produire lorsque qu'elle commence avec une entrée $s = 100$ ou $s = 200$. Il existe cependant des cas où le programme ne va jamais atteindre 100. Il est facile pour un être humain de vérifier que sitôt que le nombre s atteint la valeur 1, le programme se contentera de répéter 142142... sans jamais s'arrêter. En mathématiques, cette suite de nombres porte d'ailleurs un nom : on l'appelle la suite de Syracuse. Si vous voulez gagner de l'argent, et de la notoriété en tant que mathématicien, il vous suffit de prouver que ce programme va toujours atteindre la valeur 1 ou la valeur 100, après un nombre fini d'étapes, quelle que soit la valeur initiale de s . Il s'agit en effet d'un problème irrésolu en théorie des nombres.

D'un point de vue formel, le problème de l'arrêt, identifié par Alan Turing, consiste donc à construire une machine de Turing *Halt* qui soit capable d'analyser la sémantique d'un programme. Il doit saisir ce que le programme cherche à réaliser et utiliser ses connaissances mathématiques pour prédire l'issue de la démarche. $Halt(T, s)$ prend deux arguments : le nom de la machine T à analyser et l'entrée s de cette machine. T et s sont donc tous les deux des chaînes de caractères. Par commodité, on suppose que deux machines de Turing portant le même nom T ont toujours le même code. Pour prouver que *Halt* n'existe pas, une façon de procéder consiste à supposer que cette machine existe et à montrer que cela conduit à une contradiction. On appelle cette stratégie de preuve la démonstration par l'absurde — *reductio ad absurdum* en latin.

Plutôt que d'entrer dans les détails techniques de la preuve de l'impossibilité de *Halt*, j'ai imaginé une parabole humoristique. Je pense que celle-ci devrait vous aider à vous forger une intuition préliminaire du problème. Écoutez seulement : il y a fort longtemps, dans une vallée oubliée, existait un Maître magicien du nom de Tigran Ulan (un anagramme d'Alan Turing). Une fois par jour, il réunissait ses disciples sur la colline et leur demandait, après avoir prononcé un mot à voix haute, de faire le vide dans leur esprit. Le maître était conscient de la difficulté de cet exercice, surtout lorsque le

mot dit à voix haute est celui de « Bella » — une femme du village voisin qui attirait sur elle tous les fantasmes. Il n'y avait pas de limite de temps alloué à cet exercice, si bien qu'après une heure, il était difficile de dire, de ceux qui n'avaient pas encore réussi à faire le vide dans leur esprit, s'ils finiraient par y arriver un jour.

Pour pallier ce problème, le Maître disposait heureusement d'un sort de divination, nommé *Paradoxe*, qui lui permettait de prédire quels étaient ceux d'entre ses étudiants qui allaient réussir l'exercice après un temps fini et ceux qui, même en leur donnant un temps infini, n'y arriveraient jamais. Le fonctionnement de *Paradoxe* suivait le principe des vases communicants : lorsque le magicien lançait ce sort sur une personne dont l'esprit ne devait pas s'arrêter, il atteignait lui-même aussitôt un état de paix absolu ; lorsqu'au contraire il lançait le sort sur quelqu'un dont l'esprit finirait par s'arrêter après un temps suffisant, l'esprit du Maître se mettait à accélérer en répétant sans fin le nom de « Bella ». Cet effet était quelque peu fâcheux, mais le Maître s'y accommodait. Il disait souvent : « les apparences sont trompeuses, les derniers seront les premiers ! ».

Un jour, un des disciples du magicien, frustré d'être tenu en échec, vint trouver ce dernier et lui dit : « Maître, pourquoi ne faites-vous pas l'exercice avec nous ? Nous voulons savoir si vous savez, vous, résister aux charmes de Bella ! » Le Maître eut du mal à dissimuler son embarras. À force de lire dans l'esprit de ses étudiants, n'avait-il pas fini par fantasmer lui aussi l'innocente bachelette ? Heureusement, il avait une parade. « Si je lance *Paradoxe* sur moi-même, expliqua-t-il, devrais-je réussir ou échouer l'exercice ? »

Le disciple réfléchit quelques instants. Il comprit sans tarder que si le sortilège prédisait quelque chose à propos de l'esprit du Maître, il produirait l'effet inverse. Si par exemple l'esprit du Maître devait s'arrêter, *Paradoxe* aurait effet d'agiter cet esprit jusqu'à la nuit des temps. N'était-ce pas en contradiction avec la prédiction du sortilège ? D'autre part, si l'esprit du Maître ne devait pas trouver la paix, *Paradoxe* aurait pour effet de l'interrompre immédiatement. Le sortilège portait décidément bien son nom !

Dans cet exemple, les esprits des individus correspondent à des machines de Turing. Le mot « Belle » dit à voix haute est l'entrée s et l'état de cessation du mental représente la situation dans laquelle une machine de Turing s'arrête. Le sort *Paradoxe* peut donc être représenté par un programme qui a pour fonction de prédire si une machine de Turing va s'arrêter sur l'entrée s . Si c'est le cas, *Paradoxe* boucle infiniment. Sinon, *Paradoxe* s'arrête. Cette prédiction en question demande, naturellement, de faire recours à la fonction *Halt*. Le code peut donc s'écrire de la façon suivante :

```

procédure Paradoxe(x) :
  si Halt(x, x) répond oui,
    boucle infinie;
  sinon,
```

s'arrêter;

Il est important de noter la présence de l'expression $Halt(x, x)$. Cette ligne stipule que la lettre x puisse à la fois désigner un programme et à la fois l'entrée s donnée à ce programme. Cette condition est satisfaite par le fait que chaque programme peut être représenté par un nom et que deux programmes avec le même code portent toujours le même nom. À l'exception de ce détail technique, il en va de *Paradoxe* comme dans la parabole que j'ai donnée.

Il est facile de voir que *Paradoxe* adopte un comportement paradoxal lorsqu'on lui demande de fonctionner sur lui-même. Si on suppose en effet que $Paradoxe(Paradoxe)$ s'arrête, alors $Halt(Paradoxe, Paradoxe)$ répondra que oui, $Paradoxe(Paradoxe)$ s'arrête. Mais dans ce cas, *Paradoxe* doit enclencher une boucle infinie, ce qui est en contradiction avec le fait qu'il est censé s'arrêter. Inversement, si on suppose que *Paradoxe* ne s'arrête pas, alors $Halt(Paradoxe, Paradoxe)$ répondra que non, $Paradoxe(Paradoxe)$ ne s'arrête pas. Or dans ce scénario, le programme *Paradoxe* s'arrête, ce qui est en contradiction avec le fait qu'il n'était pas censé s'arrêter. La seule conclusion valide consiste à dire qu'il n'existe pas de machine *Halt* capable de dire quand une machine s'arrête.

5.3.3 Le Théorème de Rice

La deuxième remarque consiste en un corollaire direct du problème de l'arrêt. Si n'existe pas de machine capable de dire, d'un programme quelconque, s'il s'arrête sur une entrée quelconque, il est facile de voir qu'il n'existe pas non plus de machine susceptible de dire, d'un programme quelconque, s'il possède une propriété sémantique non triviale quelconque. Par *propriété non triviale*, j'entends une propriété qui ne soit pas vraie ou fausse pour toutes les machines de Turing possibles et imaginables. Par *propriété sémantique*, j'entends une propriété qui touche, non pas aux lignes de code en tant que chaînes de caractères, mais à ce que le code est capable d'effectuer. La question de savoir si le programme contient plus de dix lignes de code, ou s'il contient le mot « alors » n'est pas sémantique. À l'inverse, les questions ci-dessous constituent des propriétés sémantiques non triviales :

- La machine écrit-elle, à un moment de son processus, le symbole 0 sur la première case de sa bande de lecture ?
- Les sorties de la machine $T(s)$ correspondent-elles à la multiplication de s par 2 ?
- Les sorties de la machine $T(s)$ correspondent-elles à une fonction calculable $f(s)$ donnée ?

La raison pour laquelle ces propriétés ne peuvent pas être décidées par

un algorithme est liée à un phénomène que je vais décrire, qu'on appelle la semi-décidabilité. Pour vérifier qu'une machine écrit, à un moment de son processus, le symbole 0 sur la première case de son ruban, on pourrait se contenter de regarder cette case et de voir si, à un moment donné, le symbole zéro apparaît. Cette méthode est efficace pour dire qu'un programme écrit effectivement un 0 sur cette case. Mais lorsqu'au contraire rien ne se passe après, mettons, une heure d'analyse, il n'est pas évident de décider si la machine ne le fera jamais ou s'il suffit d'attendre plus longtemps.

Nous sommes, à ce titre, placés dans la même situation que les premiers hommes : voyant le soleil se coucher chaque soir, ils se demandaient s'il pouvait arriver que le soleil ne se lève pas au matin suivant ? En l'absence d'une loi de la nature concernant le déplacement des astres, cette question est indécidable. Il est d'ailleurs tout aussi incertain, comme le signalait déjà David Hume, que les lois de la nature ne rencontrent pas d'exception inattendue. On ne saura jamais, en effet, s'il ne suffirait pas d'attendre un matin de plus pour observer, enfin, l'absence du soleil. En termes techniques, on dira qu'une propriété est semi-décidable s'il est possible de vérifier les situations où cette propriété est rencontrée, mais non pas de déterminer les situations où la propriété n'est pas rencontrée.

On pourrait penser qu'il est possible de contourner cette difficulté en construisant un algorithme suffisamment intelligent ? Cet algorithme ne se contenterait pas de regarder ce qui se passe sur une case donnée, mais étudierait la structure même du programme. C'est là qu'intervient le problème de l'arrêt. Si le problème de l'arrêt n'était pas indécidable, on pourrait utiliser *Halt* pour déterminer si une machine de Turing $T(s)$ exécute une certaine fonction $f(s)$, comme par exemple la multiplication de s par deux. Il suffirait de déterminer d'abord si la machine s'arrête ou non. Dans le cas où elle ne s'arrête pas, elle ne calcule pas $f(s)$. Dans le cas où elle s'arrête, il suffit de vérifier que les symboles écrits sur la bande de lecture à l'arrêt du programme correspondent bien à $f(s)$. Il s'ensuit qu'il est impossible de construire un algorithme général permettant d'analyser et de déterminer la sémantique d'un programme quelconque. Pour contourner cette difficulté, on est contraint de :

- recourir à des raisonnements au cas par cas ;
- Procéder à des approximations ;
- Restreindre l'analyse sémantique à un sous-ensemble de programmes clefs, en faisant par exemple des hypothèses sur la structure de ces programmes.

La moralité du théorème de Rice, c'est qu'il est intrinsèquement difficile d'analyser la sémantique d'un programme. Les machines de Turing ne semblent pas très adroites lorsqu'il s'agit de comprendre la signification de leurs lignes de codes — ou celle d'une autre machine. Ce sont des exécuteurs

aveugles, qui manipulent des symboles, qui les assemblent, les divisent, les transforment, sans avoir forcément une vision générale, une auto-inspection, sur les démarches qu'elles sont en train de réaliser. Ce sont des bras mécaniques, des câbles, des processus séquentiels, dépourvus d'entendement au sens le plus exigeant du terme. Peut-être n'avait-on jamais été en position de dire, plus encore qu'aujourd'hui, que les machines fonctionnent de façon *automatique*, avec toute la désinence d'impuissance et de subtil dédain exprimée par ce mot ?

5.3.4 Oracles et cybernétique

Les limitations imposées par le théorème de Rice s'appliquent à toutes les machines de Turing, mais pas nécessairement à toutes les machines possibles et imaginables. Il serait tentant de se demander si l'on peut construire une extension de la machine de Turing qui permette de répondre à des questions indécidables, telles que celles se rapportant à la sémantique d'un programme ?

Cette extension que j'envisage n'est pas facile à trouver. Alan Turing a en effet montré que la plupart des adjonctions que l'on réalise sur une machine de Turing n'augmentaient pas sa puissance de calcul. Si l'on crée par exemple une machine dotée de plusieurs têtes de lecture ou de plusieurs bandes infinies, cette machine ne peut rien faire de plus qu'une machine de Turing classique. Ce que j'entends par là, c'est que toute fonction calculable par une machine de Turing étendue T' peut être calculée par une machine de Turing classique T , quoiqu'il lui faille souvent un peu plus de temps d'exécution. Il est donc possible d'améliorer le temps de calcul nécessaire à la résolution de questions décidables, mais on ne peut pas répondre à des questions indécidables.

Une façon de contourner ces limitations consiste à imaginer un Oracle capable de résoudre "magiquement" n'importe quelle question binaire, et ce en une seule étape de calcul. À l'instar de l'antique Pythie, ou de la Sybille des Romains, cet Oracle est en mesure de répondre par « oui » ou par non « non » sans risque de se tromper. Vous pouvez lui donner des questions décidables, du genre « la somme des angles d'un triangle fait-elle 180 degrés en géométrie euclidienne ? », mais vous pouvez aussi lui demander : « Une machine T quelconque s'arrête-t-elle sur une entrée s quelconque ? ». L'astuce derrière cette idée de l'Oracle — qui porte décidément bien son nom — c'est qu'on s'épargne la peine d'en décrire le fonctionnement. L'Oracle ne donne jamais de justifications à ses réponses. Il est entouré d'une brume, d'une aura de mystère, digne de cette phrase de Plutarque : « La Pythie s'exprime : beaucoup nient, d'aucuns l'ignorent, certains déduisent^[20] ».

Par souci de précision, notons que les Oracles ne sont pas forcément

20. PLUTARQUE, *Dialogues pythiques*.

tous capables de répondre à toutes les questions. Dans certains contextes, il est possible de restreindre les réponses formulées par un Oracle à certaines classes de questions précises. Par exemple, un Oracle peut être conçu pour dire si une certaine entrée s donnée sur une machine T appartient à l'ensemble des entrées qui s'arrête en un temps fini. Cet Oracle est donc en mesure de résoudre le problème de l'arrêt, mais pas forcément les autres problèmes. On peut aussi construire un Oracle permettant de déterminer si une machine T donnée implémente effectivement la division euclidienne ; ou autrement dit, si les réponses de cette machine correspondent au résultat attendu de la division euclidienne.

Ce qui est proprement novateur avec l'introduction de l'Oracle, c'est qu'on ne s'intéresse plus au fonctionnement interne de calcul, mais uniquement aux entrées et aux sorties. L'Oracle peut être comparé à une « boîte noire » mystérieuse (une boîte de Pandore ?) qui prend une question d'entrée et qui fournit, en sortie, la réponse attendue. Cette idée de « boîte noire » opaque s'oppose à l'idée d'une « boîte blanche » totalement transparente. Cette terminologie est introduite pour la première fois par le mathématicien Norbert Wiener, dans un article visant à modéliser la communication des machines et des animaux (en ce compris l'être humain^[21]) :

Les termes "boîte noire" et "boîte blanche" sont des expressions pratiques et figuratives dont l'usage est assez imprécis. J'entendrai par boîte noire un dispositif, tel qu'un réseau à quatre bornes avec deux bornes d'entrée et deux de sortie, qui effectue une opération définie sur le potentiel d'entrée actuel et passé, mais pour lequel nous n'avons pas nécessairement d'information sur la structure par laquelle cette opération est réalisée. En revanche, une boîte blanche sera un réseau similaire dans lequel nous avons intégré la relation entre les potentiels d'entrée et de sortie conformément à un plan structurel précis, destiné à garantir une relation d'entrée-sortie préalablement déterminée.

La réponse fournie par une boîte noire peut être introduite dans une boîte blanche ou dans une autre boîte noire. En connectant des boîtes les unes aux autres, on obtient un système d'information solidaire, où chaque étape de calcul conduit à l'étape suivante. Un exemple humoristique de cette approche peut être trouvé dans le diagramme de Marguerite Souchon, popularisé sur Twitter, pour déterminer par quel livre de Dostoïevski commencer votre lecture de la littérature russe^[22]. Dans ce diagramme, que j'ai essayé de reproduire dans la figure 5.3, chaque boîte répond par « oui » ou par « non »

21. NORBERT WIENER, *Cybernetics : Or Control and Communication in the Animal and the Machine*, Préface, page xi. (Footnote 1), 1948. La traduction est de moi.

22. Marguerite Souchon est professeur de russe et autrice du livre intitulé *Le Dieu de Dostoïevski*, paru en 2021. Le 6 octobre 2021, elle a publié sur Twitter un diagramme devenu célèbre.

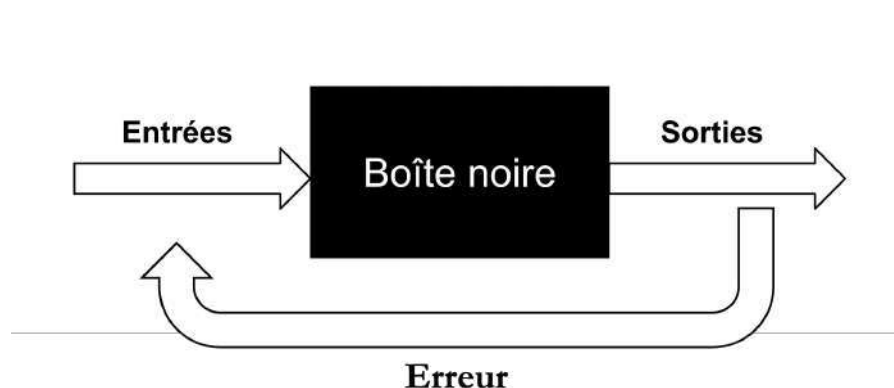


FIGURE 5.4 – Schéma représentant schématiquement le paradigme de la calculabilité. Au lieu de définir la raison comme une méthode déductive partant des axiomes pour aboutir à des conclusions, on peut la représenter comme une boîte noire qui, à chaque question introduite en entrée, fournit une réponse en sortie.

sur lui sont si nombreuses qu'elles menacent, sans cesse, de le déstabiliser. Lorsque l'engin penche sur la gauche, il faut activer la turbine droite pour le rétablir ; lorsqu'il penche sur la droite, on active la turbine opposée, et ainsi de suite. Ce système peut être représenté par un ensemble de boîtes noires, qui prend en entrée les paramètres initiaux du drone (inclinaison, vitesse, position) et qui prédit en sortie la trajectoire future du drone (inclinaison, vitesse, position). Cette prédiction, à son tour, est comparée à l'état que l'on souhaite atteindre (celui d'un vol stationnaire). L'écart entre l'état idéal et l'état prédit est appelé *l'erreur*. Elle est réinjectée dans le système initial de façon à corriger la trajectoire à chaque nouvelle étape de calcul.

Remarquons que, dans l'exemple du drone en vol stationnaire, les boîtes noires ne répondent pas à des questions binaires avec des réponses sous la forme « oui » ou « non ». Elles répondent à des questions numériques du type : à quelle vitesse activer la turbine ? selon quel angle incliner l'appareil ? etc. Les travaux du mathématicien Wiener peuvent être vus comme une extension du concept de l'Oracle. Ils sont souvent considérés comme étant à l'origine des systèmes cybernétiques, définis comme des systèmes d'information complexes.

Il est indifférent, dans le paradigme de la cybernétique, que la boîte noire fournisse des réponses en se basant sur une fonction calculable ou en recourant à de simples associations statistiques. Tout ce qui importe, c'est que la réponse soit bonne, ou le plus proche possible d'une bonne réponse. Face à des questions qui sont soit trop complexes pour être calculées, soit intrinsèquement incalculables, il est tentant de recourir à des modèles statistiques donnant des réponses appropriées par simple imitation des cas connus.

Les modèles statistiques actuellement les plus performants, ce sont les réseaux de neurones. À première analyse, ces derniers se comportent exactement comme des boîtes noires. Ils donnent des réponses sans fournir de justification, en imitant le plus fidèlement possible les cas connus. On ne comprend pas clairement la fonction liant l'entrée à la sortie, et on n'est pas même sûr que cette fonction existe sous une forme manipulable. Ce que je veux dire par là, c'est qu'il n'est peut-être pas possible d'exprimer une relation générale, simple, élégante et compréhensible par un esprit humain, entre les entrées et les sorties. Tout se passe donc comme si les réseaux de neurones étaient les Oracles de l'informatique moderne.

Dans la prochaine partie de cette thèse, j'aimerais suggérer que les théories de la démontrabilité et de la calculabilité définissent deux façons bien différentes d'envisager la programmation. Dans le premier cas, on s'intéresse à des systèmes de preuves, basés sur des axiomes et des lois logiques, qui procèdent étapes par étapes ; de tels systèmes peuvent être représentés par des boîtes blanches. Dans le second cas, on s'intéresse à des systèmes de boîtes noires interconnectées, basés sur des données statistiques, et dont on ne peut pas prédire le comportement *a priori*. Ces deux points de vue, complémentaires, sont les deux visages de l'intelligence artificielle. Ils forment ce que j'appellerai la programmation *top-down* et la programmation *bottom-up*.

L'abandon de la théorie du jugement de Kant a donné naissance au formalisme de Hilbert. Dans ce mouvement de pensée, le jugement n'est plus affaire d'intuition, mais de logique pure.

Plus précisément, la raison est définie comme une certaine aptitude à manipuler des symboles d'après certaines règles, sans égard à la signification donnée à ces symboles. Un bon exemple est la résolution d'une équation mathématique, qui demande de manipuler des lettres algébriques en oubliant momentanément ce que ces lettres désignent.

Le formalisme de Hilbert participe de l'idée que la pensée puisse être automatisée par une machine. Les premiers ordinateurs ne sont d'ailleurs rien d'autre que des machines à manipuler des symboles.

L'automatisation se heurte toutefois à certains problèmes : d'une part, la logique présente des paradoxes et, d'autre part, il existe des problèmes indécidables. Ces problèmes ont pour conséquence que certaines vérités logiques et mathématiques ne peuvent pas être découvertes à l'aide de règles top-down univoques, mais uniquement pas des méthodes bottom-up.

Dans le prochain chapitre, je vais montrer que cette impasse a conduit à deux paradigmes de programmation bien différents.

Chapitre 6

Approche Top Down : Les Systèmes Experts

Le rêve de Hilbert était de trouver une méthode capable de déterminer, par un ensemble de règles purement mécaniques, le vrai du faux. Si une telle méthode existait, on pourrait *de facto* se passer du jugement humain. Il n'y aurait non seulement rien que la machine ne puisse pas faire, en termes de recherche scientifique, mais cette machine serait du reste plus fiable, plus précise, plus rapide que nous autres. Or nous avons vu que ni l'approche de Frege, basée sur les axiomes, ni celle d'Alan Turing, basée sur les machines portant son nom, ne permettent de totaliser ou d'automatiser les savoirs. Autrement dit, il existe toujours des vérités indémontrables, aussi bien par le moyen d'une démonstration mathématiques que d'un calcul informatique, basé sur la machine de Turing.

L'échec du rêve d'automatisation de Hilbert est une bonne nouvelle pour notre démarche. Cela signifie que l'oubli de l'Entendement qui caractérise la naissance de l'ordinateur, et qui s'est produit comme on l'a vu dès l'irruption des géométries non euclidiennes, présente des limites théoriques. Il ne semble pas possible d'expliquer le savoir humain en se fondant uniquement sur la démontrabilité ou sur la calculabilité. Tout laisse à penser que, entre ces deux paradigmes, existe un autre ingrédient fondamental, qui a été délaissé par l'informatique à ses débuts. S'agit-il de l'intuition ? S'agit-il d'une faculté de juger ?

Le paradigme de la démontrabilité s'appuie uniquement sur la *syntaxe*. Elle cherche à formaliser les disciplines mathématiques, à commencer par l'arithmétique, en les réduisant à un ensemble d'axiomes. Une proposition du type « $5 + 7 = 12$ », au sein de ce système, sera considérée vraie si et seulement s'il existe un chemin inférentiel allant des axiomes à la proposition. Ce chemin n'est pas unique, mais il est bien déterminé : on peut en tracer les étapes. En revanche, on ne sait pas à quoi ressemblerait la machine qui suivrait ces étapes. Il y a certes un mode d'emploi abstrait, mais pas de

modèle concret.

Par ailleurs, et il s'agit là d'une autre difficulté encore, les théorèmes d'incomplétude de Gödel, comme je l'ai indiqué, suggèrent que la grande majorité des propositions sont indémontrables. Le savoir axiomatique, obtenu de façon purement analytique, semble dès lors constituer l'exception plutôt que la règle. Cet îlot de raison, de clarté, de lumière, ce paradis des Idées louangé par Hilbert, et avant lui Platon, ne constitue qu'un refuge minuscule dans le régime froid, insondable, énigmatique, de la connaissance humaine. Comment pouvons-nous, sur cette base fragile, espérer embrasser du regard la vaste étendue des connaissances ?

Le paradigme de la calculabilité n'est pas davantage épargné par la critique. Il échappe à prime abord aux théorèmes de Gödel. Il prend le point de vue de la sémantique. Il ne s'intéresse pas tant à comprendre les règles logiques du raisonnement déductif qu'à décrire, quant à ses principes, le fonctionnement d'une machine automatique. Cette dernière est munie d'une tête de lecture, d'un ruban infini, de symboles, d'un mécanisme aussi huilée qu'une authentique machine de Charles Babbage^[1]. Il devient possible de la créer de toute pièce, de lui soumettre des problèmes, de l'utiliser pour encoder ou décoder des messages^[2]. Ce que l'on gagne sur le plan concret, on le perd toutefois en traçabilité. L'on aimerait pouvoir prédire parfaitement tout ce qu'une machine va faire si on inscrit une entrée x donnée sur son ruban, mais le problème de l'arrêt et le théorème de Rice font obstacle à ce rêve de transparence. Il faut admettre que la machine peut adopter des comportements émergents, que des réponses peuvent apparaître qui n'étaient pas prédites de façon préalable.

Il existe une tension épistémologique entre la syntaxique et la sémantique qui n'est pas sans rappeler celle que j'avais décrite, dans l'introduction de cette thèse, entre l'analyse et la synthèse. La syntaxe s'intéresse en effet à comprendre les lois du raisonnement. On cherche les axiomes, on s'interroge sur les forces et les limitations de la logique. Tout est donné de façon transparente. Il n'y a pas un seul résultat qui ne puisse pas, ultimement, être justifié par des preuves formelles. La raison humaine, dans ce paradigme, s'apparente à un grand dictionnaire juridique, muni de lois, de clauses et d'une certaine jurisprudence. Il n'y a pas de mystère, pas de surprise, pas d'approximation, pas d'ambiguïté.

La sémantique, pour sa part, s'intéresse moins à comprendre les lois du

1. La machine de Charles Babbage, aussi appelée la machine analytique, est une machine mécanique visant à exécuter n'importe quel programme de calcul. Elle ne fut jamais construite, mais le mathématicien Babbage lui a donné une description minutieuse. Il s'agit du premier prototype conceptuel de l'ordinateur.

2. C'est d'ailleurs bien ce que fera l'équipe d'Alan Turing pour décoder Enigma, la machine utilisée par les Allemands durant la Seconde Guerre mondiale. Ils se basent pour ce faire sur les travaux postérieurs de Marian Rejewski et sur des machines électromagnétiques surnommées « bomba kryptologiczna ».



raisonnement qu'à simuler une machine capable de raisonner. Cette machine prend des données et les assemble selon des règles organiques, pour obtenir quelque chose de nouveau. Certes, son fonctionnement n'est ni contingent, ni énigmatique. Mais pour savoir ce que va donner une machine de Turing, il est nécessaire de lancer le programme. Si on connaissait d'avance le résultat, en quoi aurions-nous besoin du programme ? La machine ne fait jamais rien que des actions déterministes : elle suit des étapes qui, une à une, sont prédictibles. Elle s'apparente néanmoins davantage à une expérience de laboratoire qu'à une démonstration mathématique. On lui donne une entrée, on regarde le résultat. On modifie l'entrée et on constate en quoi cela influe sur la réponse. On peut la représenter comme une boîte noire, qui associe à chaque input un output déterminé.

Lorsque ces deux paradigmes sont utilisés pour concevoir des intelligences artificielles, l'on aboutit à des résultats bien différents. L'approche syntaxique et axiomatique conçoit la cognition de façon logico-déductive. Les programmes informatiques sont donc rédigés sous forme d'un ensemble d'instructions, qui doivent être lues ligne après ligne. Après avoir suivi scrupuleusement le mode d'emploi, la machine est censée fournir la réponse correcte. Bien que l'ordinateur offre une efficacité et une précision hors de toute commune mesure, les premières formes d'algorithme attestées en archéologie remontent à l'antique Babylone. Il s'agit de tablettes d'argiles reprenant la recette d'une ratatouille de poireaux³ :

La viande n'est pas utilisée.
 Vous préparez de l'eau.
 Vous ajoutez de la graisse.
 Vous ajoutez
 du poireau égyptien (kurrat),
 de la coriandre,
 du sel selon votre goût,

3. Ces instructions sont attestées sur une tablette d'argile remontant à 1 700 ans avant Jésus-Christ. Bien que les inscriptions soient endommagées, deux recettes sont encore visibles. Celle que j'ai écrite ci-dessous (je l'ai traduite depuis l'anglais) ainsi qu'une recette de rôti d'agneau. Voir *Food in Ancient Mesopotamia : Cooking the Yale Babylonian Culinary Recipes* paru dans *Ancient Mesopotamia Speaks : Highlights of the Yale Babylonian Collection*.

```

        du poireau,
        de l'ail.
    Vous écrasez du levain sec,
    Vous le tamisez
    Vous le saupoudrez sur la marmite avant de la retirer.

```

Un algorithme plus sophistiqué et plus moderne, consisterait à déterminer si un nombre entier n donné est un nombre premier. Pour ce faire, on tente de le diviser par les entiers compris entre 2 et $n - 1$. Si on y arrive, on arrête la procédure : ce n'est pas un nombre premier. Si, après avoir opéré ces divisions, on n'est arrivé à aucun résultat, on conclut que n est un nombre premier. En formalisant cette procédure, cela donne la chose suivante :

```

Fonction prem(n):
    Si n = 0 ou n = 1 alors retourner Faux
    Pour chaque i compris entre 2 et n-1 :
        Si i divise n, retourner Faux
    Sinon retourner Vrai

```

Dans les exemples ci-dessus, les règles sont décrites de façon *top-down*. La tâche à réaliser — ou l'aptitude que l'on cherche à modéliser — est définie par avance. On part d'un certain problème, comme vérifier certaines propriétés sur les nombres, identifier un objet géométrique au sein d'un environnement virtuel ou modéliser un phénomène hydrodynamique, et à partir de ce point, le design consiste à définir un algorithme de résolution. Tout se passe à la façon d'une recette de cuisine : c'est en vue de réaliser un certain but que des procédés sont mis au point. Le succès de ces procédés reposera sur un certain nombre d'hypothèses ; et tant que l'on reste confiné au sein de ce contexte donné, les résultats seront garantis. Cette méthode, qui est aussi la première à avoir vu le jour de façon convaincante, est aussi appelée la programmation orientée sur les buts, ou *goal oriented*.

L'approche sémantique et synthétique, à contre-pied de ces recettes de cuisine, cherche à construire des corrélations significatives entre les inputs donnés à la machine et ses outputs. Au lieu de lui dire explicitement les étapes à réaliser pour obtenir la recette, on laisse un ensemble de paramètres à sa discrétion : la température de l'eau, la quantité de chaque ingrédient, la vigueur avec laquelle le mélange est réalisé, etc. Tout l'art de la programmation consiste alors à enseigner à la machine comment choisir au mieux ces paramètres. La stratégie la plus simple repose sur la Loi sempiternelle du bâton et de la carotte. On demande à la machine de créer des milliers de repas avec des paramètres aléatoires et de conserver ceux qui sont au plus proche de la recette voulue. En procédant ainsi à plusieurs reprises, en créant entre chaque tentative des variations et en gardant les plus concluantes, on peut espérer converger vers la recette idéale.

Cette approche peut être qualifiée de *bottom-up*. C'est-à-dire qu'on ne prescrit pas par avance les tâches à réaliser. On se contente de soumettre à la machine un ensemble de données à ingurgiter, à décortiquer, à analyser. À l'issue de cette indigestion, l'intelligence artificielle détermine elle-même les règles à suivre pour obtenir un résultat. Bien sûr, il revient à l'être humain de fixer les lois d'apprentissage, en disant à la machine comment lire les données et les utiliser pour régenter ses comportements. L'illustration la plus éloquente de cette approche est celle d'un ordinateur tel que Hall, dans l'Odyssée de l'espace. En plus d'être un véritable agent conversationnel, cette machine est capable de déterminer sa propre marche opératoire. Il ne se contente pas simplement, à façon d'un ordinateur de bord des vieux airbus, d'annoncer sans fin les mêmes protocoles. Il se montre au contraire capable de s'adapter aux nouveaux contextes, de déterminer ses alliés, ses ennemis et, le cas échéant, de faire des sacrifices.

Ce qui est intéressant avec ces deux paradigmes, c'est qu'ils tentent d'approcher l'intelligence humaine par des voies différentes, quoique complémentaires. Tandis que la stratégie top-down cherche à comprendre la cognition en la dotant des règles logico-déductives données de façon apriorique, la stratégie bottom-up cherche à l'imiter au moyen d'associations statistiques aposterioriques. Cette division entre le raisonnement pur et l'expérience pure n'est pas sans rappeler la division que Platon, et Kant après lui, opéraient entre la Sensibilité et la Raison. En procédant de façon logico-déductive, on est assuré de ne jamais produire une réponse qui ne soit pas strictement issue des principes de départ, mais on ne sait rien créer de nouveau (la recette babylonienne n'évoluera jamais). Dans l'approche statistique, il y a une créativité intrinsèque (on peut découvrir de nouvelles recettes), mais on n'est jamais à l'abri d'un résultat indésirable (on peut mettre le feu à la cuisine). C'est pourquoi j'aimerais poser, à titre d'hypothèse, que les méthodes statistiques de type bottom-up permettent d'imiter le jugement synthétique a posteriori, tandis que les déductions top-down se réduisent à des jugements analytiques *a priori*.

Hypothèse 3. *Le paradigme top-down est uniquement capable de produire des jugements analytiques, tandis que le paradigme bottom-up est uniquement capable de produire des jugements synthétiques a posteriori.*

Il est important de noter que je parle bien des paradigmes de programmation top-down et bottom-up et non pas des paradigmes théoriques de l'informatique que sont les théories de la démontrabilité et de la calculabilité. Bien que la syntaxe de l'axiomatisation ait encouragé la stratégie de programmation top-down, il n'est pas interdit de l'utiliser pour décrire le fonctionnement d'un réseau de neurones typiquement bottom-up. Les lois statistiques peuvent en effet être axiomatisées; il est possible de raisonner sur ces systèmes de façon logico-déductive. Inversement, lorsque je dis de la sémantique de la machine de Turing qu'elle participe de la stratégie

de programmation bottom-up, cela n'exclut pas la possibilité d'utiliser une machine de Turing pour prouver des axiomes. Après tout, les algorithmes top-down tournent bien sur un ordinateur. Fondamentalement, la machine de Turing et les systèmes d'axiomes sont deux représentations différentes d'un même concept : celui d'un processus automatique de calcul. Ils ne sont donc pas en eux-mêmes top-down ou bottom-up. Néanmoins, ils ont inspiré des façons de voir l'informatique, et de construire ses outils, qui sont quant à elles diamétralement opposées.

Les chapitres qui suivent cherchent à justifier l'hypothèse que je viens de poser en décrivant de la façon la plus exacte possible les outils techniques dont ces deux paradigmes sont constitués. Il s'agira à la fois de comprendre ce qu'il y a en eux d'analytique ou de synthétique, mais également de repérer par avance certaines zones d'ombre. En critiquant ces outils, en jetant de la lumière sur les problèmes où ils échouent à apporter des jugements convaincants, j'espère tailler à coup de becs et d'ongles un espace pour une troisième espèce de jugement, occupant le milieu entre la Sensibilité et la Raison. On verra que ce jugement n'aura plus exactement la même forme que les jugements synthétiques *a priori* kantien.

Avant de commencer l'analyse que je me suis donnée pour tâche, il peut être utile de poser quelques questions préliminaires : lequel des deux paradigmes, top-down et bottom-up, est le plus utilisé aujourd'hui ? Quant ont-ils vu le jour exactement ? Quels rapports entretiennent-ils entre eux ? Au vu de l'histoire récente de l'intelligence artificielle, il est plausible de penser qu'on ait assisté au glissement progressif allant d'une approche majoritairement top-down (auquel ce chapitre est dédié) vers une inclination de plus en plus marquée pour le bottom-up (qui fera l'objet du prochain chapitre).

Jusqu'au début des années 2000, il n'était pas rare de dire que les ingénieurs, loin de s'affairer à la conception d'un droïde aux apparences humanoïdes capable de parler et de se comporter à l'image d'un être humain, s'efforçaient bien plutôt de mettre au point des algorithmes dédiés à la réalisation de tâches bien circonscrites : jouer aux échecs, optimiser des flux financiers, assurer la logistique, etc. On répétait l'idée que l'informatique, sérieusement envisagée, devait se consacrer au perfectionnement de l'intelligence artificielle faible, et non à la mythologie d'un robot pensant⁴. Or, aujourd'hui, force est de constater qu'on n'a jamais été aussi près de concevoir les premières formes d'intelligence artificielle véritablement générales. L'exemple de ChatGPT jouant aux échecs, répondant à des questions de

4. Pour rappel, l'intelligence artificielle faible désigne un logiciel capable de résoudre certaines tâches spécifiques, comme jouer aux échecs, tandis que l'intelligence artificielle forte concerne une machine dotée de toutes les capacités cognitives humaines, ce qui la rend flexible, autonome et adaptable. Le lecteur peut se référer au chapitre pour lire la définition de l'intelligence artificielle forte d'après François Chollet. Notez que certains auteurs ajoutent encore un troisième niveau, la super intelligence artificielle, qui aurait une conscience et dépasserait largement les aptitudes humaines.

physique théorique ou échangeant, comme par jeu, le scalpel de laboratoire contre la plume de poète, a si bien défrayé la chronique qu'il est aujourd'hui envisageable de se consacrer à la mise au point d'une architecture algorithmique indépendante du contexte.

Ce changement de paradigme est d'autant plus surprenant qu'il semblerait que, dans le chef de l'être humain, l'on assiste à un phénomène en tout point asymétrique. Pour nous autres citoyens, il semblerait que le monde du travail demande de nous une plus grande spécialisation. Il est loin le temps où les universités formaient, sous la houlette de Léonard de Vinci, des savants universels. Aujourd'hui, il en va de la condition d'existence du travailleur et du chercheur que de s'emparer, par la force de ses compétences, d'un lopin de terre dans les territoires immenses et bien cartographiés de la connaissance. Nous sommes ainsi en détention d'outils dont les origines (les données embrassées) et les finalités (les tâches possibles) échappent de plus en plus à l'étroitesse de notre vision (l'hyperspécialisation). Est-ce à dire que nous n'avons plus aucune chance de comprendre les machines auxquelles nous avons donné le jour ?

« Là où croit le danger, écrivait Hölderlin, croît aussi ce qui sauve. » C'est de notre intelligence collective, et de cette vaste somme d'expériences et de points de vue dont notre société est composée, que sont nées les machines auto-adaptatives ; n'est-ce pas là également que les solutions doivent être envisagées ? C'est grâce à l'Internet, ce puits inaltérable de savoirs interconnectés, que sont collectés quotidiennement les big data. L'Internet est ce qui permet à l'intelligence collective d'être, pour ainsi dire, synthétisée dans nos algorithmes. Leurs comportements, leurs biais, leurs systèmes de références, sont à l'image statistique de nos opinions, nos croyances, nos discours, nos agissements. Nous façonnons un monde qui est, en tout point, à l'image de notre propre monde intérieur. Nous le faisons d'une façon telle, toutefois, qu'il ne semble plus nous appartenir en propre, mais être la possession exclusive du genre humain en son entier⁵. Dès lors, ne serait-ce pas également grâce à cette intelligence collective que nous pourrions continuer à comprendre et à orienter, dans un futur immédiat, les progrès de l'intelligence artificielle ?

Pour donner un cadre historique et méthodologique solide à ces réflexions (dont je touche un mot dans les chapitres [16.2.2](#) et [16.2.2](#)), il me faut caractériser le passage allant de l'approche top-down à l'approche bottom-up.

5. À ce titre, il est un fait notable que la propriété intellectuelle ait été profondément remise en question par l'apparition de Chat GPT. Tout se passe comme si l'ancien régime de conception des savoirs, d'après lequel l'individu est, par la force de son esprit, capable d'insuffler son intelligence propre au monde extérieur, était ébranlé par un nouveau règne. Dans ce dernier, les créations sont collectives et, pour ainsi dire, anonymes. Lorsque Dally, ou tout autre générateur d'images automatiques, produit le portrait d'une vieille femme, celui-ci rassemble, en une image synthétique, la production de tous les photographes à avoir publié leurs créations sur l'Internet. Dans ce visage ridé, fané, brillant tout entier de la force des générations, ne se trouverait-il pas un peu du sourire de votre grand-mère ?

Un ouvrier, avant que de s'engager à l'exécution d'une menuiserie, s'avise de comprendre les outils qu'il manipule, ainsi que l'impact que ceux-ci peuvent avoir sur son travail. Il est fort à parier que son rapport à la construction soit influencé par l'apparition, dans son atelier, de machines industrielles ; et, sitôt que son ciseau laisse la place aux découpeuses automatiques, les principes et la discipline qu'il se donne pour atteindre le résultat optimal ont bien des chances de changer.

Dans le même ordre d'idées, le philosophe qui se penche sur l'intelligence artificielle est tenu de comprendre en quoi l'évolution de l'outil visant à modéliser et comprendre la pensée humaine a, en retour, influencé l'exercice de cette pensée. C'est pourquoi, dans les pages qui vont suivre, je vais décrire brièvement le glissement allant du paradigme top-down, bien circonscrit et maîtrisé, au paradigme bottom-up. Plus particulièrement, je serai attentif aux principes épistémologiques, ou à la conception implicite de l'esprit humain, qui sont véhiculés par chacune de ces approches.

6.1 The General Problem Solver

En 1967 eut lieu, entre les murs du Massachusetts Institute of Technology (MIT), une partie d'échecs d'un genre particulier. L'un des protagonistes, un éminent professeur de philosophie du nom de Hubert Dreyfus, fit face au logiciel Mac Hack VI. On imagine aisément la scène. D'un côté, une chaise blanche rapiécée, de ces chaises au dossier élimé, bossue par le poids des années d'études, de siestes et de recherches fébriles qui ont traversé l'institut universitaire. En face se tient un écran noir strié de formes brillantes. L'humain est assis dans une pose qui n'aurait rien à envier au Penseur de Rodin. Du haut de sa chaise, il défie la machine imperturbable.

Si ce n'était le ronronnement grésillant des circuits électroniques, on pourrait entendre une mouche voler. Autour des deux adversaires défilent une série de visages blêmes, aux lunettes tordues, aux cravates ligaturées. Les sourcils se froncent, perplexes. De la sueur coule sur les tempes imberbes. Les lèvres laissent échapper un sourire d'ironie au moment où la reine blanche, après une brillante incursion, se fait prendre à revers.

Un souffle de soulagement traverse la partie de la salle occupée par les informaticiens. Il faut les comprendre : Hubert Dreyfus avait eu l'audace de déclarer, quelque temps plus tôt, que nul ordinateur ne parviendrait jamais à vaincre un enfant de dix ans. Certes, l'assertion était provocante ; mais cette boutade, toute audacieuse qu'elle fût, n'avait d'égale que cette autre prophétie : « D'ici une décennie, avait-on crié en finançant le secteur de la recherche à coups de millions de dollars, un ordinateur sera en mesure de vaincre un champion d'échecs ! »

Le philosophe Dreyfus n'était pas le champion tant attendu. Tout au plus faisait-il figure d'amateur zélé. Un peu gauche. Brillant de temps en

temps. Trop enthousiaste, peut-être, pour convaincre. Mais il n'était pas non plus un enfant — ni prodige, ni débutant. En toute logique, on aurait dû s'attendre à un match nul. Ce jour-là, pourtant, la machine l'emporta.

Parmi les spectateurs en liesse se tenait un homme de quarante ans. Son front plus haut que les autres, ses lunettes plus épaisses, contrastaient avec son fin sourire. Son nom est Herbert Simon. C'est lui qui, dix ans plus tôt, avait tenu l'affirmation que la machine battrait le champion. Il est conscient, certes, d'avoir été un peu trop assertif, mais l'issue du match constitue tout de même, en faveur de l'optimisme informatique, un beau baroud d'honneur. « C'était une partie magnifique, commentera-t-il plus tard. Le suspense était intense entre ces deux stratèges aux étincelles de génie, aux plans diaboliques... avec les drames et les désastres qui accompagnent souvent ce genre de rencontres⁶. »

Loin d'être simplement un match opposant l'homme à la machine, cette partie d'échecs du MIT symbolise, plus fondamentalement, la tentation antinomique de deux opinions, deux Muses, aux rêves grandiloquents. D'un côté se tient le philosophe pénétré de l'idée que la machine ne pourra pas, en l'état actuel, imiter la pensée humaine dans toutes ses dimensions. De l'autre côté se trouve l'informaticien pétri de l'ambition de concevoir une machine capable de résoudre toutes sortes de problèmes humains. D'un côté, donc, la Muse de chair et de nuages, au visage digne des plus illustres tableaux romantiques ; de l'autre, la Muse de cuivre et de silicium, tout droit ressuscitée du film de Metropolis (ou des mains de Frankenstein).

L'affrontement sur l'échiquier, en tant que tel, n'avait rien de nouveau dans l'histoire de l'intelligence artificielle. Depuis la construction du célèbre Turc mécanique, qui avait affronté Napoléon, il y avait eu bien des rencontres entre l'homme et la machine. Certaines d'entre elles, certes, étaient pipées — un maître d'échecs ne se cachait-il pas derrière les rouages bien huilés du Turc ? Mais ce qui est proprement inaugural avec la partie contre Mac Hack VI, c'est que, pour la première fois de l'histoire de l'humanité, pouvait être esquissée une description opérationnelle de ce à quoi ressemblerait une machine pensante.

Vers la fin des années cinquante, en effet, Herbert Simon avait développé, de concert avec ses collègues Allen Newell et John Shaw, un logiciel intitulé le General Problem Solver (GPS). En dépit de son nom rondet, le programme visait à convaincre les plus sceptiques de la capacité d'un ordinateur à solutionner n'importe quel problème. Il suffisait d'introduire, sous forme de propositions logiques, une description des objets à manipuler — les axiomes de la géométrie euclidienne, les axiomes de la logique prédicative, etc. — pour que le logiciel entreprenne, au moyen de la recherche opérationnelle, une déduction de la meilleure solution admissible. Une des spécificités de ce

6. Traduit depuis *Oral History of Richard Greenblatt*, interviewed by GARDNER HENDRIE.

programme, c'est qu'il dissociait l'encodage des méthodes de résolution de l'encodage du problème proprement dit. Les auteurs de l'article résument ce fonctionnement à l'aide du graphique **???**

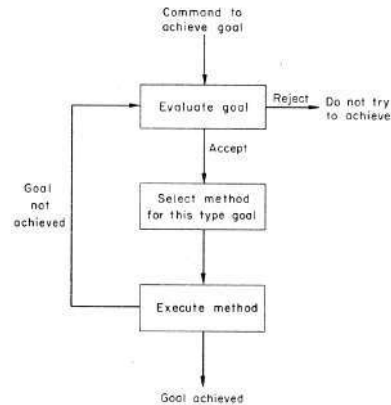


Fig. 1—Executive organization of GPS

Le principe fondamental du *General Problem Solver* consiste à prendre pour données les axiomes d'une théorie et à produire un maximum de déductions. En raison de la complexité calculatoire, le nombre de possibilités explose bien rapidement, ce qui demande d'affiner quelque peu l'approche par force brute. Il faut déterminer en particulier quelles sont les données pertinentes à considérer en amont du processus et quelles sont celles intéressantes à produire en aval de ce processus. Pour contourner ces difficultés calculatoires, trois ingrédients fondamentaux furent introduits par Newell, Simon et Shaw :

- **Arbres de possibilités** : La machine parcourt méthodiquement un arbre de possibilités. Aux racines de cet arbre se trouvent les données disponibles (axiomes de la théorie et relations connues). Les ramifications, quant à elles, représentent chacune une déduction nouvelle. Cet arbre de possibilités doit être parcouru jusqu'à rencontrer la proposition que l'on cherche à démontrer (ou son impossibilité). Cette façon de procéder est au fondement de la recherche opérationnelle, un champ très important en algorithmique.
- **Heuristiques** : Pour éviter de devoir parcourir aveuglément l'ensemble des possibilités, Newell et Simon font usage d'heuristiques. Ces règles intuitives, déterminées préalablement par l'être humain, sont fondées sur la connaissance précise du problème. Un expert en chirurgie ne va pas nécessairement envisager toutes les hypothèses. Son

7. A. NEWELL, J. C. SHAW, H. A. SIMON, *Report on a general problem-solving program*, 30 décembre 1958.

expérience lui permet d'exclure, *a priori*, et de façon souvent inconsciente, certaines possibilités. À vrai dire, le raisonnement conscient, chez l'être humain, n'interviendra que sur un très petit nombre de cas possibles perdus dans l'immensité des possibilités.

- **Listes :** Afin de gagner en temps d'exécution, l'information traitée par le programme est représentée sous forme de listes de type $A \rightarrow B \rightarrow C$, où les éléments se suivent de façon séquentielle. Cette approche s'apparente à la notation du lambda calcul utilisée par Church. Elle est au fondement du langage de programmation LISP, encore exploité de nos jours.

La machine est en outre dotée de deux espaces de stockage : la mémoire de travail et la mémoire à long terme. La première contient les opérations en cours de calcul, tandis que la seconde contient l'ensemble des faits connus par le programme, soit parce qu'ils appartiennent aux prémisses de base, soit parce qu'ils ont été démontrés. Cette façon de représenter la mémoire au sein de la machine n'est pas sans évoquer l'analogie de ma tablette de cire de Platon. Tout se passe comme si les données étaient des symboles manipulables, que l'on peut imprimer quelque part dans les ressorts de la machine, et rappeler à volonté. Pour ce faire, elle se fonde sur deux importantes règles de déduction :

1. **Chaîne ascendante :** le programme part des axiomes et tente de les combiner stratégiquement pour prouver la formule évaluée par le programme (dans quel cas elle est acceptée) ou pour prouver sa négation (dans quel cas elle est refusée).
2. **Chaîne descendante :** le programme part de la formule et tente de la simplifier pour la réduire à des axiomes de la théorie (dans quel cas la formule est acceptée) ou à leur négation (dans quel cas elle est refusée). Il suffit qu'un seul axiome soit contredit pour disqualifier la formule.

Les chaînes ascendantes et descendantes sont comme les deux termes d'une égalité $A = B$. Pour prouver cette relation, vous pouvez soit développer le membre de droite, soit le membre de gauche, soit les deux successivement. Dans les Systèmes Experts, l'intelligence est conçue, purement et simplement, comme la manipulation de symboles dans un espace de mémoire numérique. Tout objet de connaissance doit être introduit sous la forme d'un nombre dans la machine, de sorte que celle-ci opère, dans un second temps, le calcul visant à la production d'un autre nombre — la réponse.

Cette approche, quoique prometteuse, s'apparente à une calculatrice sophistiquée. Elle emprunte son intuition épistémologique au rêve de Hilbert

lorsqu'il disait, des mathématiques, qu'elles n'étaient rien d'autre que l'ensemble des signes tangibles dérivables, par un être humain consciencieux, sur une feuille de papier vierge. Dans le General Problem Solver, les signes manipulés n'ont pas d'interprétation *a priori*. Il échoit à celui qui introduit les données, et qui recueille les réponses de la machine, de leur donner une signification, en disant de telle variable qu'elle désigne, par exemple, une température, un coefficient de friction, etc.

Dans les pages qui suivent, je vais donner une définition plus formelle du système expert. Cela va demander, dans un premier temps, de représenter le raisonnement déductif ascendant et descendant au moyen d'un langage mathématique. Il s'agira ensuite de comprendre comment résoudre, au sein de ce langage, des questions ouvertes du type « vrai » ou « faux ».

6.2 Logique propositionnelle

Dans la mesure où la machine manipule des symboles tels que des lettres, (A, B, C...) des nombres (1, 2, 3...) ou des opérateurs fonctionnels spécifiques ($\rightarrow$, $\vee$...), il semble important de se demander si ces symboles peuvent représenter la pensée humaine ? S'il était en effet avéré que la pensée était autre chose qu'une pure et simple manipulation de signes abstraits, l'on tiendrait là un argument essentiel pour tracer une ligne de démarcation entre l'être humain et l'ordinateur, du moins l'ordinateur tel qu'on le connaît aujourd'hui.

Il se trouve que cette préoccupation, celle qui consiste à représenter les idées avec des signes, était déjà présente dans le chef des anciens philosophes. Des figures telles qu'Aristote ou Chrysippe ont largement contribué à développer et à formaliser l'étude du raisonnement, en montrant quels rapports, quelles successions peuvent ou ne peuvent pas légitimement être observées entre les idées pour être tenues pour valides.

La logique consiste fondamentalement en une abstraction consistant à réduire la pensée à un ensemble de termes isolés les uns des autres. À ce titre, la logique présuppose que les processus cognitifs peuvent être mis à plat sur le papier et être représentés de façon linéaire, chaque moment de la pensée conduisant nécessairement au moment suivant. Cette façon d'exprimer le contenu de nos réflexions contraste largement avec notre expérience subjective.

La logique, en imposant la linéarité du raisonnement, capture un aspect fondamental de l'esprit humain : les liens de nécessité entre les idées. Lorsque l'on entre dans le champ de la logique, on ne s'intéresse plus à la pensée telle qu'elle est vécue dans l'expérience subjective, mais bien au résultat objectif de la pensée. Il y a des associations qui semblent spontanément autorisées entre les idées, et d'autres non. Il est par exemple autorisé de dire que si Socrate est un homme, et si tous les hommes sont mortels, alors

Socrate est mortel. En revanche, on trouvera suspicieux de conclure à partir de là que tous les hommes sont Socrate. Cela vient de ce qu'il existe une relation d'appartenance entre les concepts rationnels que nous manipulons. Socrate appartient à l'ensemble des humains, et ces humains eux-mêmes appartiennent à l'ensemble des êtres mortels.

Les syllogismes ne font rien d'autre que formuler, sous la forme la plus générale possible, l'existence de cette relation d'appartenance qui semble présente dans l'esprit humain. Par « forme générale », j'entends une expression qui peut valoir quels que soient les termes envisagés. Elle doit pouvoir s'appliquer à Socrate et aux hommes mortels, mais également aux carrés, aux rectangles, aux quadrilatères. C'est pourquoi l'abstraction de la logique, non contente de linéariser les raisonnements, recourt également à des symboles. On écrira par exemple :

$$\text{Si } A \rightarrow B \text{ et si } B \rightarrow C, \text{ alors } A \rightarrow C$$

Dans cette expression, la flèche $\rightarrow$ désigne l'implication et peut se lire « si le terme de gauche est vrai, alors le terme de droite est vrai », ou encore « chaque fois que le terme de gauche est vrai, on observe également que le terme de droite est vrai ». Il est aisé de voir qu'en remplaçant A par « être Socrate », B par « être un homme » et C par « être mortel », on retrouve le syllogisme de notre exemple : si être Socrate implique d'être un homme et qu'être un homme implique d'être mortel, alors être Socrate implique d'être mortel. Le raisonnement tient toujours en posant $A =$ être un carré, $B =$ être un rectangle et $C =$ être un quadrilatère.

L'expression ci-dessus utilise le vocabulaire et la grammaire de la logique dite *propositionnelle*. Cette logique, très simple à formaliser, n'utilise qu'un très petit nombre d'ingrédients, mais elle peut exprimer un assez grand nombre de pensées. D'un point de vue technique, la logique propositionnelle peut être définie comme un langage comportant un alphabet, une grammaire et une sémantique de façon à construire des formules déductibles au sein d'un système axiomatique.

6.2.1 Alphabet

L'alphabet est l'ensemble de tous les symboles utilisés dans le langage. Dans le cadre de la logique propositionnelle, on ne trouve que deux types de symbole : des éléments propositionnels en nombre infini dénombrables (noté $p_1, p_2, \dots, q_1, q_2, \dots, r_1, r_2, \dots$) ainsi que des connecteurs logiques ($\vee, \wedge, \neg, \rightarrow, \leftrightarrow$) et des parenthèses (« $\)$ » et « $(\)$ »). On verra plus loin ce que ces symboles signifient et comment ils doivent être interprétés. Pour l'instant, contentons-nous de dire que :

- La négation $\neg$ est placée devant un symbole propositionnel pour exprimer l'idée que ce qu'il désigne est faux.

- La disjonction $\vee$ correspond à un « ou » : elle est placée entre deux symboles propositionnels pour exprimer l'idée qu'au moins l'un des deux symboles désigne quelque chose de vrai. Il n'est pas exclu que les deux soient vrais.
- La conjonction $\wedge$ correspond à un « et » : elle est placée entre deux symboles propositionnels pour exprimer l'idée que les deux symboles désignent quelque chose de vrai.
- L'implication $\rightarrow$ est placée entre deux symboles propositionnels pour exprimer l'idée que si le premier désigne quelque chose de vrai, alors le second le fait également.
- La bi-implication $\leftrightarrow$ est placée entre deux symboles propositionnels pour exprimer l'idée que ces symboles désignent tous les deux des choses vraies ou qu'ils désignent tous les deux des choses fausses. On dira qu'ils ont la même *valeur de vérité*.

Remarquons qu'on trouve parfois l'usage d'opérateurs non standards, comme la barre de Scheffer ($|$) ou la disjonction exclusive (noté $\oplus$ ou w). Il est toutefois démontré que ces opérateurs sont redondants. Seul un sous-ensemble d'entre eux suffit pour écrire toutes les phrases possibles⁸.

6.2.2 Grammaire

La grammaire stipule comment il est possible d'écrire des phrases en se servant de ce vocabulaire. Dans la logique propositionnelle, on peut écrire n'importe quelle phrase consistant en une séquence de symboles propositionnels de type $p_1, p_2, \dots$ séparés les uns des autres par des opérateurs logiques. La phrase $p_1 \rightarrow p_2$ est une phrase valide, tandis que $p_1 p_2 \rightarrow p_3$ n'est pas valide, car les symboles p_1 et p_2 ne sont pas séparés par un opérateur. Nous ne savons dès lors pas comment les lire. Notons que lorsque plusieurs opérateurs interviennent dans une même phrase, il est important d'utiliser des parenthèses pour distinguer quelle opérateur s'applique à quel symbole. On procède ici de façon semblable à ce qui se fait en arithmétique. Le but est d'empêcher toute forme d'ambiguïté de notation.

Si l'on dénote par L l'ensemble des expressions bien fondées de la logique propositionnelle, on peut formaliser ce qui vient d'être dit ci-dessus de la manière suivante :

L1 Tous les symboles propositionnels sont des éléments de L .

L2 Si $A \in L$ et $B \in L$, alors

L2a $\neg A \in L$

L2b $A \vee B \in L$

L2c $A \wedge B \in L$

8. On peut par exemple se contenter de prendre $\vee$ et $\wedge$. Alternativement, on peut se contenter d'utiliser uniquement la barre de Scheffer $|$.

- L2d $A \rightarrow B \in L$
 L2e $A \leftrightarrow B \in L$
 L3 Rien d'autre n'est un élément de L .

6.2.3 Sémantique

Ces phrases dont nous avons donné les règles d'écriture n'auraient aucun intérêt si une sémantique ne leur était pas attachée. Par sémantique, j'entends deux choses : (1) une certaine façon de lire les symboles propositionnels, de les interpréter, de façon à faire coïncider la phrase avec un raisonnement en français et (2) une manière de déterminer si une formule utilisant ces symboles est vraie ou fausse.

Le point (1) exige d'associer les symboles propositionnels $p_1, p_2, \dots, q_1, q_2, \dots$ à des propositions du langage. Le point (2) consiste à se demander quelles sont les conditions pour qu'une formule utilisant ces symboles, du type $q_1 \rightarrow q_2$, soit elle-même vraie ou fausse.

Il est important de noter que la logique propositionnelle est une logique bivaluée. Elle stipule qu'il n'existe que deux valeurs de vérité possibles : le vrai ou le faux. Cette conviction qu'une chose est soit vraie, soit fausse, est ce qu'on appelle le principe du tiers exclu. Aristote s'en fait le défenseur lorsqu'il écrit :

Il n'est pas possible qu'il y ait aucun intermédiaire à l'intérieur de la contradiction, mais il est nécessaire d'affirmer ou de nier une seule chose d'une seule chose^[9].

Il précise toutefois qu'une chose n'est vraie ou fausse que dans la pensée, c'est-à-dire dans la façon dont une idée est associée ou non à une autre :

Le faux ni le vrai ne sont point dans les choses, comme, par exemple, si le bien était le vrai, et le mal, le faux. Ils n'existent que dans la pensée; encore, les notions simples, la conception des pures essences, ne produisent-elles rien de semblable dans la pensée^[10].

Par convention, on associera chaque élément propositionnel $q_1, q_2, \dots$ à une valeur de vérité dépendant de l'interprétation de ce terme. La valeur de vérité est égale à 1 si l'élément propositionnel est vrai et égale à 0 s'il est faux.

À titre d'exemple, posons $p_1 =$ « Celui ou celle qui lit cette phrase est un être humain » et $p_2 =$ « Celui ou celle qui lit cette phrase est mortel ». Ces propositions peuvent être tantôt vraies, tantôt fausses, en fonction du

9. ARISTOTE, *Métaphysique*, IV, 7, 1011b23-24.

10. ARISTOTE, *Métaphysique*, III, 25.

contexte. Lorsque vous lisez cette phrase, on a $p_1 = \text{vraie}$ et $p_2 = \text{faux}$, mais si un dieu de la mythologie la lisait, on aurait $p_1 = \text{faux}$ et $p_2 = \text{vrai}$.

On se servira ensuite de règles pour déterminer si une phrase telle que $q_1 \wedge q_2$ doit être jugée vraie ou fausse. Ces règles sont déterminées en suivant la signification intuitive des opérateurs logiques : dans la mesure où le symbole $\wedge$ représente un "et" logique, la phrase $q_1 \wedge q_2$ sera vraie si et seulement si q_1 et q_2 sont tous les deux vrais ; la phrase $q_1 \vee q_2$, quant à elle, est vraie si et seulement si un des deux éléments propositionnels est vrai, etc.

Ces règles sont résumées dans le tableau ci-dessous, appelé aussi *table de vérité* :

A	B	$\neg A$	$A \vee B$	$A \wedge B$	$A \rightarrow B$	$A \leftrightarrow B$
0	0	1	0	0	1	1
0	1	1	1	0	1	0
1	0	0	1	0	0	0
1	1	0	1	1	1	1

Dans l'exemple que j'ai donné pour q_1 et q_2 , la phrase $q_1 \rightarrow q_2$ se lit « Si celui ou celle qui lit cette phrase s'appelle Socrate, alors celui ou celle qui lit cette phrase est de genre masculin ». Intuitivement, cette phrase ne peut être fausse que si la personne qui la lit est un être humain et que cette personne est immortelle. Dans notre monde, cela n'est pas possible, mais cela ne veut pas dire que la phrase soit vraie de toute éternité, et ce pour deux raisons.

La première, c'est qu'on peut envisager des mondes, certes mythologiques, où un être humain est immortel. Dans la mythologie hindoue, les Chiranjivis sont des êtres humains ayant reçu le don de l'immortalité. Le plus connu d'entre eux est Hanuman.

La seconde raison, c'est que, tout en restant dans notre monde connu, il est possible de donner aux propositions q_1 et q_2 une autre interprétation que celle de notre exemple. On pourrait dire $q_1 =$ « celui qui lit cette phrase est un être humain » et $q_2 =$ « celui qui lit cette phrase est de genre masculin ». Sous cette nouvelle interprétation, il est facile de trouver des cas où la formule $q_1 \rightarrow q_2$ est fausse. Il suffit d'envisager qu'une femme lise cette phrase.

Une question intéressante en logique consiste dès lors à se demander s'il existe des phrases dont la valeur de vérité ne dépend pas de l'interprétation des éléments dont elle est constituée. On peut, à ce titre, distinguer trois cas de figure :

- **Tautologies** : Certaines phrases sont toujours valides quelle que soit l'interprétation des éléments propositionnels. À titre d'exemple, la phrase $q_1 \vee \neg q_1$ est une tautologie qui dit : « soit q_1 est vraie, soit

q_1 est fausse ». C'est une loi logique. On dira que la formule est valide.

- **Contradictions :** Certaines phrases ne sont jamais valides, quelle que soit l'interprétation des éléments propositionnels. À titre d'exemple, la phrase $q_1 \wedge \neg q_1$ est une contradiction qui exprime en substance : « q_1 est vraie et fausse à la fois ». On dira que la formule est invalide.
- **Contingences :** Certaines phrases peuvent être vraies ou fausses, selon l'interprétation des termes. Par exemple $q_1 \wedge q_2$ est vraie si et seulement si q_1 et q_2 sont vrais. Cela peut être obtenu en prenant par exemple $q_1 =$ « le ciel est bleu » et $q_2 =$ « le soleil est un astre ».

Les tautologies sont particulièrement intéressantes, car elles peuvent être utilisées pour réaliser des raisonnements. Le fait qu'elles ne soient jamais fausses permet tantôt de simplifier certaines expressions, tantôt de montrer que deux relations que l'on pensait fort différentes sont, en réalité, équivalentes, comme on va le voir dans la section suivante.

6.2.4 Formules

Il est important de remarquer qu'il est possible d'écrire des propositions logiques prenant pour variables, non pas les éléments propositionnels $p_1, p_2, \dots$, mais des formules $A, B, \dots$. Quelle est au juste la différence entre $p_1 \vee p_2$ et $A \vee B$? Dans le premier cas la phrase doit être interprétée « soit on a p_1 , soit on a p_2 », où p_1 et p_2 peuvent représenter n'importe quelle idée. Ils peuvent être associés à la sémantique $p_1 =$ « rouge » et $p_2 =$ « noir », tout comme ils pourraient être associés à la sémantique $p_1 =$ « petit » et $p_2 =$ « grand ».

Les lettres A et B figurant dans le second cas de figure, quant à elles, représentent n'importe quelle formule propositionnelle. On peut donc substituer l'expression $p_1 \rightarrow p_2$ à la lettre A et $p_1 \vee (p_2 \vee p_3)$ à la lettre B . Dans le cas particulier où A vaut p_1 et B vaut p_2 , on retrouve l'expression $p_1 \vee p_2$, mais il s'agit d'un cas particulier. Pour le dire autrement, l'usage de lettres A et B est bien plus général que les symboles propositionnels. Si les symboles propositionnels sont utilisés pour exprimer des relations unissant des idées, des éléments individuels quelconque de la pensée ; les lettres majuscules, quant à elles, sont mobilisées pour exprimer des relations unissant des raisonnements les uns aux autres.

6.2.5 Axiomes

Le raisonnement humain est caractérisé, entre autres propriétés, par la capacité à réaliser des déductions. Ces dernières consistent à partir de certaines propositions logiques et à les assembler pour déduire des vérités plus

générales. L'exemple typique de déduction est donné dans la syllogistique d'Aristote : si l'on sait que Socrate est un homme ($s \rightarrow H$) et que les hommes sont mortels ($H \rightarrow M$), on en déduit que Socrate est mortel ($S \rightarrow M$). En des termes modernes, on parlera de transitivité. Ce procédé de pensée met en relief l'existence de lois logiques, décrétant quelles sont les déductions valides et quelles sont celles qui sont trompeuses.

Dans la logique des propositions, ces lois peuvent être écrites sous la forme d'un certain nombre d'axiomes. Ceux-ci prennent comme variables des formules A, B, C du langage propositionnel et nous dictent comment elles peuvent être associées les unes aux autres. Autrement dit, les axiomes sont des phrases qui sont toujours vraies, quelle que soit l'interprétation de ces termes A, B et C .

- AS1 $A \rightarrow (B \rightarrow A)$
- AS2 $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$
- AS3 $((A \rightarrow B) \rightarrow A) \rightarrow A$
- AS4 $(A \wedge B) \rightarrow A$
- AS5 $(A \wedge B) \rightarrow B$
- AS6 $A \rightarrow (B \rightarrow (A \wedge B))$
- AS7 $A \rightarrow (A \vee B)$
- AS8 $B \rightarrow (A \vee B)$
- AS9 $(A \rightarrow C) \rightarrow ((B \rightarrow C) \rightarrow ((A \vee B) \rightarrow C))$
- AS10 $(A \leftrightarrow B) \rightarrow (A \rightarrow B)$
- AS11 $(A \leftrightarrow B) \rightarrow (B \rightarrow A)$
- AS12 $(A \rightarrow B) \rightarrow ((B \rightarrow A) \rightarrow (A \leftrightarrow B))$
- AS13 $\neg\neg A \rightarrow A$
- AS14 $(A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$

Bien que ces axiomes puissent surprendre à première vue, ils correspondent à des méthodes de raisonnement tout à fait usuelles. L'axiome 4 par exemple nous dit que si on sait que $A \wedge B$ est vrai, alors on peut en déduire que A est vrai. Cela se note :

$$A \wedge B \vdash A$$

Le symbole $\vdash$ intervenant dans cette expression indique qu'il existe une déduction logique partant d'une phrase appelée la prémisse ($A \wedge B$) pour écrire une nouvelle phrase appelée la conclusion (A)¹¹. Ces déductions sont définies comme un ensemble d'étapes, ou de lignes de symboles, qui s'appuient sur les phrases vraies connues pour obtenir de nouvelles phrases vraies.

11. S'il existe plusieurs phrases à droite du symbole $\vdash$, il faut au moins qu'une des phrases soit vraie. Ce cas est cependant rarement rencontré et constitue une sorte de convention.

Notons que les axiomes que j'ai présentés ci-dessus sont également munis d'une règle d'inférence ad hoc intitulée le Modus Ponens . Celle-ci consiste à dire que si on a A et $A \rightarrow B$, alors on peut écrire B . Cette règle s'écrit :

$$A, A \rightarrow B \vdash B$$

On est dès lors en mesure de donner une définition formelle du concept de déduction logique comme étant un raisonnement composé d'un certain nombre d'étapes, représenté sous forme de lignes, et justifié à partir des axiomes.

Proposition 3. *Une déduction axiomatique de B à partir des prémisses $A_1, A_2, \dots, A_n$ est une liste de lignes telles que chaque ligne soit :*

- une prémisses, ou
- un axiome de la logique classique propositionnelle (AS1–AS14), ou
- un résultat D de la règle MODUS PONENS : si C est sur la ligne i dans la liste et $C \rightarrow D$ est sur la ligne j de la liste, D peut être dérivé sur une ligne k ($k > i$ et $k > j$) de la liste

et où B est l'élément final de la liste.

Notons que j'ai utilisé ici le terme déduction axiomatique, signifiant que la déduction ne fait intervenir que les axiomes. Dans la pratique d'un système expert, il est commode d'utiliser, à certaines étapes du raisonnement, des déductions qui ont déjà été prouvées précédemment.

Pour illustrer cette approche, supposons d'abord que nous soyons en mesure de prouver le Modus Tollens de façon axiomatique. Le Modus Tollens, parfois aussi appelé la contraposée, est une relation fort connue en logique, qui dit que si A implique B et que B est faux, alors on ne peut pas avoir A . Si en effet A était vrai, on aurait B . En logique, cela s'écrit :

$$\neg B, A \rightarrow B \vdash \neg A$$

Une preuve axiomatique peut être donnée en utilisant les axiomes AS1 et AS14 :

1	$\neg B$	PREMISSE
2	$A \rightarrow B$	PREMISSE
3	$\neg B \rightarrow (A \rightarrow \neg B)$	AS1
4	$A \rightarrow \neg B$	MODUS PONENS (lignes 1 et 3)
5	$(A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$	AS14
6	$(A \rightarrow \neg B) \rightarrow \neg A$	MODUS PONENS (lignes 2 et 5)
7	$\neg A$	MODUS PONENS (lignes 4 et 6)

Il est important de remarquer que les symboles de l'axiome AS1 l'on retrouve à la ligne 3 ne sont pas identiques aux symboles que j'ai utilisés en introduisant cet axiome. L'axiome S1 était en effet défini de la façon suivante :

$$A \rightarrow (B \rightarrow A)$$

Or dans la preuve, j'ai utilisé :

$$\neg B \rightarrow (A \rightarrow \neg B)$$

Pour comprendre cette substitution, il faut se rappeler que les symboles A et B intervenant dans l'écriture de l'axiome AS1 désignent n'importe quelle formule. Il est donc tout à fait permis de choisir la formule $\neg B$ à la place de A et la formule A à la place de B . Un procédé analogue s'applique dans le cas de l'axiome AS14 mobilisé à la ligne 5.

Maintenant que la preuve axiomatique du Modus Tollens est chose faite, il est commode d'utiliser le Modus Tollens au sein d'une déduction sans devoir à nouveau passer par les axiomes. Si, par exemple, je fais une déduction dans laquelle j'ai déjà établi, lors des étapes antérieures, qu'on a $\neg B$ et $A \rightarrow B$, je peux aussitôt écrire $\neg A$ à la prochaine ligne de mon raisonnement. De ce que je viens de dire, il en découle une nouvelle définition.

Proposition 4. *Une déduction (pas nécessairement axiomatique) de B à partir des prémisses $A_1, A_2, \dots, A_n$ est une liste de lignes telle que chaque ligne soit :*

- *une ligne autorisée par la déduction axiomatique, ou*
- *le résultat D d'une règle $C_1, C_2, \dots, C_n \vdash D$, où les prémisses $C_1, C_2, \dots, C_n$ figurent toutes parmi les lignes antérieures de la liste*

et où B est l'élément final de la liste.

6.3 Les Systèmes Experts

La logique propositionnelle qui a été présentée peut être utilisée par un Système Expert pour réaliser des déductions. Un tel système présente alors trois types de composants :

1. **Les axiomes de la logique :** les lois fondamentales du raisonnement humain, représentées dans la forme la plus simple et la plus abstraite.
2. **Les lois de la théorie :** les lois fondamentales du champ disciplinaire dans lequel s'exerce le Système Expert. S'il s'agit d'une machine ayant pour vocation d'établir des diagnostics à propos de tumeurs cancéreuses, il va falloir la doter d'un certain nombre d'hy-

pothèses concernant la façon dont les observations cliniques (encéphalogrammes, rayons X...) peuvent être utilisées pour établir des conclusions valides (présence d'une tumeur bénigne, maligne..). Ces hypothèses sont écrites sous la forme des lois logiques et ajoutées aux axiomes de la logique proprement dite.

3. **Les données du problème :** Les observations cliniques disponibles sur le patient doivent être encodées dans le Système Expert sous forme d'éléments propositionnels de types $q_1, q_2, \dots$, auxquels on associe des valeurs de vérité. Si q_1 représente la présence d'une protubérance et qu'il n'y a pas de protubérance, on donnera $q_2 = \text{faux}$ à la machine. S'il y a une protubérance, on donnera $q_3 = \text{vrai}$. Si on ne sait pas, on ajoutera dans la machine la formule $q_1 \vee q_2$.

On comprendra qu'il est crucial que le langage logique utilisé soit suffisamment précis pour pouvoir encoder les lois de la théorie et les données du problème. Or le langage propositionnel présente certaines lacunes. Tout d'abord, il ne permet pas d'exprimer d'autres valeurs de vérité que vrai ou faux. Bien que cette distinction paraisse suffisante à première vue, elle ne fait pas de nuance entre ce qui est « probablement faux » et ce qui l'est « certainement ». L'on pourrait donc être tentée de donner des valeurs de vérité, non pas strictement égale à 1 ou égale à 0, mais comprise sur l'intervalle $[0;1]$. Une telle approche permet d'évoluer vers une logique probabiliste.

Un autre problème vient de ce que la logique propositionnelle ne supporte pas la contradiction. Si le système contient deux prémisses paradoxales, comme p_1 et $\neg p_1$, il devient possible de démontrer tout et son contraire. Il est en effet possible de s'appuyer sur p_1 pour démontrer certains résultats, puis de s'appuyer sur $\neg p_1$ pour démontrer le résultat contraire. Cette propriété du système est appelée le principe d'explosion et s'écrit :

$$A, \neg A \vdash B$$

Une preuve axiomatique peut être donnée par :

1	A	PREMISSE
2	$A \rightarrow (\neg B \rightarrow A)$	AS1
3	$\neg B \rightarrow A$	MODUS P. (lignes 1 et 2)
4	$\neg A$	PREM
5	$\neg A \rightarrow (\neg B \rightarrow \neg A)$	AS1
6	$\neg B \rightarrow \neg A$	MODUS P. (lignes 1 et 2)
7	$(\neg B \rightarrow A) \rightarrow ((\neg B \rightarrow \neg A) \rightarrow \neg \neg B)$	AS14
8	$(\neg B \rightarrow \neg A) \rightarrow \neg \neg B$	MODUS P. (lignes 3 et 7)
9	$\neg \neg B$	MODUS P. (lignes 6 et 9)
10	$\neg \neg B \rightarrow B$	AS13
11	B	MODUS P. (lignes 9 et 10)

Le principe d'explosion est particulièrement gênant pour les systèmes experts, puisque les bases de données utilisées pour permettre à la machine de raisonner sont susceptibles de contenir des contradictions. Envisageons par exemple un système visant à diagnostiquer la présence d'une tumeur cancéreuse. Il peut être utile de donner à cette machine toutes les observations cliniques établies à propos d'un patient. Celles-ci peuvent avoir été établies à différents moments par différentes personnes. Elles dépendent des appareils de mesures, des compétences du spécialiste, de ses hypothèses préalables et peuvent en outre contenir des erreurs. Il est donc probable que des contradictions apparaissent entre les données. On aimerait toutefois que, si le système expert conclut qu'il y a une tumeur, ce ne soit pas en s'appuyant sur le principe d'explosion, sans quoi on pourrait observer des raisonnements absurdes du type : "Le patient a 37 de fièvre et 36 de fièvre. Il s'agit là d'une contradiction. J'en déduis donc qu'il a le cancer. »

Une limitation soulevée par la logique propositionnelle, dans le contexte qui nous intéresse, vient de l'impossibilité de distinguer la possibilité de la nécessité. On aimerait pouvoir distinguer des phrases du type « l'apparition de tel symptôme rend impossible l'existence d'une tumeur » avec des phrases du type « l'apparition de tel symptôme rend possible l'existence d'une tumeur ». Or en logique des propositions, l'implication $\rightarrow$ désigne toujours une relation de nécessité, sans exception. Il faut donc raffiner la logique propositionnelle en ajoutant un symbole $\Box$ devant les propriétés qui sont nécessaires et le symbole $\Diamond$ devant celles qui ne sont que possibles.

Un dernier problème de la logique propositionnelle est la pauvreté expressive. On ne peut pas faire de distinctions, au moyen des éléments propositionnels entre le sujet et le prédicat. Pour chaque sujet i qui nous intéresse au sein d'un ensemble d'objets et pour chaque propriété j au sein d'un ensemble de propriétés, il faut créer un terme p_{ij} qui est vrai si le sujet i possède la propriété j . Ce langage ne permet toutefois pas de construire des relations complexes entre les propriétés et leurs variables. Il ne permet pas non plus d'envisager des fonctions transformant les variables, comme par exemples les opérations mathématiques de l'addition, de la soustraction, de

la multiplication. Il est dès lors impossible de poser des questions du type : « le produit de deux nombre quelconque est-il divisible par deux ? » ou encore « existe-t-il un nombre plus grand que la somme de deux nombres quelconques ? ». Pour introduire ce degré de raffinement, il peut être utile de se tourner vers la logique prédicative qui introduit des variables, des prédicats, des fonctions, ainsi que des opérateurs du type " pour tout " ($\forall$) et il existe ($\exists$).

Dans la suite de ce chapitre, et en dépit des défauts que je viens de soulever, je me concentrerai uniquement sur la logique propositionnelle. La raison en est les conclusions générales obtenues en logique propositionnelle, en ce qui concerne les systèmes experts, se transposent aux logiques plus avancées. Or mon but ici n'est pas d'établir de façon exhaustive tout ce qui se fait en matière d'algorithmique, mais uniquement d'introduire le lecteur aux notions essentielles de l'approche Top Down. Le lecteur constatera du reste que les outils dont je parle dans cet état de l'art seront utilisés ou discutés dans les parties ultérieures de la thèse.

6.4 Forme normale conjonctive

Lorsque les enfants jouent à « ni oui ni non », ils établissent des restrictions sur les mots qu'ils peuvent utiliser pour exprimer leur pensée. Cela ne retire toutefois rien à leur pouvoir d'expressivité. Tout ce que la langue permet de dire à l'aide des mots « oui » et « non » peut l'être en utilisant des synonymes (quoique l'on perde certaines nuances émotionnelles). De façon analogue, nous pourrions nous demander quels sont les plus petits ensembles d'opérateurs $\neg$, $\vee$, $\wedge$, $\rightarrow$, $\leftrightarrow$ nécessaires pour exprimer les lois de la logique propositionnelle ?

Il s'agit là d'une question importante pour mettre au point des Systèmes Experts. S'il était établi qu'il existait de nombreuses façons différentes d'exprimer les mêmes lois logiques, cela complexifierait les déductions de la machine. Elle devrait en effet prendre plus de temps pour déterminer si deux expressions sont identiques. Par ailleurs, elle risquerait de produire des relations qui ne sont que des réécritures inutiles des informations dont nous disposons déjà¹². C'est pourquoi il est judicieux de se demander s'il existe une manière unifiée d'exprimer les relations logiques ? Idéalement, cette unification doit être faite de sorte à satisfaire certains critères :

12. De manière générale, la logique propositionnelle ne permet que de déduire des tautologies les unes des autres, si bien que tout ce que l'on obtient sera toujours marqué du sceau de la trivialité. Il existe néanmoins des cas dans lesquels la déduction logique peut permettre à l'être humain de faire de la clarté sur certains problèmes. Il est toutefois légitime de penser que si Simon et Newell se fussent souvenus des enseignements de Kant, ils n'auraient pas affirmés avant tant de conviction que leur General Problem Solver — une machine spécialisée uniquement dans la résolution de jugements analytiques — fût en mesure de répondre à toutes les questions possibles.

- **Simplicité** : Cette forme unifiée doit réduire au maximum le nombre d'opérateurs logiques ;
- **Pouvoir expressif** : Elle ne doit pas restreindre l'expressivité de la machine ;
- **Calculabilité** : Elle doit accélérer au maximum le temps de calcul pris par la machine ;
- **Élégance** : Elle doit être suffisamment élégante pour permettre à un être humain de comprendre le résultat final (et idéalement les étapes intermédiaires du raisonnement).
- **Traçabilité** : Elle doit être suffisamment claire pour permettre à un être humain de comprendre les étapes intermédiaires du raisonnement.

Pour s'engager sur cette voie, il peut être utile de faire remarquer que l'implication $A \rightarrow B$ est équivalente à $\neg A \vee B$. Une simple inspection du tableau de vérité permet en effet de vérifier que l'implication $A \rightarrow B$ est vraie si et seulement si l'antécédent A est faux ou si le conséquent B est vrai.

Cette propriété n'est pas intuitive, car elle signifie que si le conséquent est trivialement vrai, on peut mettre absolument tout ce que l'on souhaite comme antécédent. Je peux par exemple prendre pour antécédent « S'il pleut et qu'il ne pleut pas » (une tautologie) et construire la phrase P « S'il pleut et qu'il ne pleut pas, alors soit je suis vivant, soit je ne suis pas vivant. » Cette phrase P sera vraie, et ce pour deux raisons.

La première raison, c'est que l'antécédent « S'il pleut et qu'il ne pleut pas » est toujours faux. Or ne peut être erronée qu'une implication où quelque chose de vrai conduit à quelque chose de faux. L'implication se contente de se demander si un événement ou un état de choses A est accompagné d'un événement ou d'un état de choses B . Si cette chose A n'a jamais lieu, on ne peut rien en dire de faux.

La deuxième raison, c'est que le conséquent « soit je suis vivant, soit je ne suis pas vivant » figurant dans la proposition P est toujours vrai. Il est donc impossible de falsifier la proposition P. Il en va de cette situation comme d'une fumée qui planerait, en tout temps et en tout lieu, sur la surface de la terre. Puisqu'elle serait omniprésente, cette fumée, on pourrait dire qu'elle accompagne le feu, mais aussi l'eau, l'air, la glace, le vent. Il sera donc tout aussi juste de dire « il n'y a pas de glace sans fumée » que de dire « il n'y a pas de feu sans fumée ».

Ces explications que j'ai exhibées montrent que l'implication $\rightarrow$ est purement formelle : elle ne tient pas compte de l'existence d'un lien substantiel entre les idées qui sont exprimées. Dans notre exemple, en l'occurrence, il n'y a pas de lien logique entre « il pleut et il ne pleut pas » et donc « soit

je suis vivant, soit je suis mort ». La seule chose qui nous intéresse est de savoir dans quel cas une de ces propositions peut être vraie tandis que l'autre est fausse, et réciproquement. Il n'est sans doute pas inapproprié de faire remonter cette vision formelle des choses à Wittgenstein, pour qui le monde est constitué, non pas de sujets et de propriétés, mais de faits atomiques qui sont connectés les uns aux autres.

1.1 The world is the totality of facts, not of things.

2.0121 Just as we cannot think of spatial objects at all apart from space, or temporal objects apart from time, so we cannot think of any object apart from the possibility of its connexion with other things.

C'est en raison de l'absence de liens matériels exprimés entre les idées que l'implication $\rightarrow$ de la logique propositionnelle peut être remplacée par les symboles $\vee$ et $\wedge$ sans affaiblir l'expressivité du langage. Je montrerai plus loin que cette simplification accélérera l'efficacité des calculs. Du reste, elle ne semble pas moins élégante à lire par un esprit humain. Il est même plus simple, à première vue, de s'exprimer uniquement avec des « et » et des « ou » qu'avec des « si... alors ».

De la même façon que $\rightarrow$ a pu être éliminée, il est aussi possible de remplacer $A \leftrightarrow B$ par $A \rightarrow B \wedge B \rightarrow A$, ce qui conduit à $(\neg A \vee B) \wedge (\neg B \vee A)$. Cette dernière expression est ce qu'on appelle une forme conjonctive normale, définie comme une conjonction ($\wedge$) de disjonctions ($\vee$). D'un point de vue technique, il est possible de montrer que toute proposition logique est réductible à une forme normale conjonctive. Pour ce faire, il convient d'introduire quelques rudiments de vocabulaire.

Définition 7. *Un littéral l est un terme propositionnel précédé, ou non, du symbole $\neg$.*

Définition 8. *Une clause en logique propositionnelle est un ensemble dénombrable de n conjonctions du type $l_1 \vee l_2 \vee \dots \vee l_n$, où $n \in \mathbb{N}$ et où $l_1, l_2, \dots, l_n$ sont des littéraux.*

Définition 9. *Une forme normale conjonctive est une conjonction dénombrables de n clauses du type $C_1 \wedge C_2 \wedge \dots \wedge C_n$, où $n \in \mathbb{N}$ et où $C_1, C_2, \dots, C_n$ sont des clauses.*

Proposition 5. *Toute formule de la logique propositionnelle peut être écrite sous une forme normale conjonctive.*

Lorsqu'une formule est écrite sous forme normale conjonctive, il suffira, pour savoir si elle est vraie, de vérifier que toutes les clauses $C_1 \wedge C_2 \wedge \dots \wedge C_n$ dont elle est constituée sont vraies. Il faut donc vérifier pour chaque clause $l_1 \vee l_2 \vee \dots \vee l_n$ s'il existe (1) une déduction allant de cette clause vers les

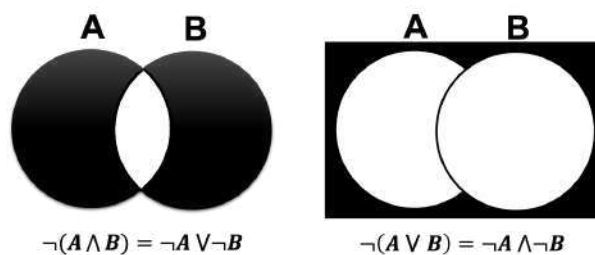


FIGURE 6.1 – Les lois de Morgan permettent d'exprimer des conjonctions comme une négations de disjonctions et des disjonctions comme des négations de conjonctions.

lois et les axiomes de la théorie (méthode descendante) ou s'il existe (2) une déduction allant des lois et des axiomes de la théorie vers cette clause (méthode ascendante).

Pour vérifier qu'une formule écrite sous forme normale conjonctive $C_1 \wedge C_2 \wedge \dots \wedge C_n$ est fausse, il faudra inversement vérifier qu'au moins une des clauses $C_1, C_2, \dots, C_n$ est fausse. Cela revient à déterminer si la négation d'une de ces clauses, notées $\neg C_1 \vee \neg C_2, \dots, \neg C_n$, est vraie. Or la négation d'une clause $l_1 \vee l_2 \vee \dots \vee l_n$ (où $l_1, l_2, \dots, l_n$ sont des littéraux) se note $\neg(l_1 \vee l_2 \vee \dots \vee l_n)$ et est équivalente à $(\neg l_1 \wedge \neg l_2 \wedge \dots \wedge \neg l_n)$. Cette transformation découle des lois de Morgan (voir graphique). Vérifier la négation d'une clause revient donc à vérifier la véracité d'une forme normale conjonctive dont chaque clause ne contient trivialement qu'un seul terme.

Il en découle de ces analyses que, pour étudier la véracité ou la fausseté d'une formule normale conjonctive $C_1 \wedge C_2 \wedge \dots \wedge C_n$, le système expert devra toujours, en dernier ressort, tenter de réduire les clauses $C_1, C_2, \dots, C_n$ (ou leur négation) aux lois de la théorie et aux axiomes. Les algorithmes de recherche peuvent donc commodément se concentrer sur la façon de manipuler des clauses, de les transformer, de les combiner en utilisant les lois de la logique propositionnelle. Il est possible de concevoir des méthodes de calculs très efficaces en suivant cette stratégie.

6.5 Problème de satisfaisabilité (SAT)

La vocation d'un système expert est, en substance, d'être capable de répondre à toute question qui lui adressée. Cette question, on l'a vu, doit être introduite dans la machine au moyen d'un langage logique et ce langage logique, en dernier ressort, peut être présenté sous une forme normale conjonctive. Il s'ensuit que le système expert doit, idéalement, être en mesure de déterminer la vérité ou la fausseté d'une expression du type :

$$\bigwedge_i^m \vee_i l_{m,i}$$

où n est le nombre de clauses, m_i est le nombre de littéraux contenus dans la i ème clause (avec $0 \leq i \leq n$) et où les $l_{m,i}$ sont des littéraux.

Nous avons vu qu'une formule logique est soit tautologique, soit contradictoire, soit contingente. J'ai indiqué dans la section précédente comment le système expert pouvait valider les propositions tautologiques et invalider les propositions contradictoires, en analysant les clauses une par une et en tentant de les rapporter aux lois de la théorie et aux axiomes de la logique. Il reste néanmoins à se demander ce qu'il se produit en présence de propositions contingentes ?

Une formule F n'est contingente que s'il existe des situations où elle est vraie et d'autres où elle est fausse. Dans le langage de la logique propositionnelle, cela consiste à dire que la chose suivante : il est possible d'assigner une valeur de vérité aux termes propositionnels $p_1, p_2, \dots$ dont la formule F est constituée de telle sorte que F soit vraie ou de telle sorte que F soit faux.

Un grand nombre de questions susceptibles d'être envisagées par un expert ne recevront pas immédiatement une réponse concluante. Il est même probable que, au plus cet expert est qualifié, au moins il sera prompt à poser une réponse définitive. Là où le novice trace un chemin direct, et souvent truffé de préjugés, entre l'observation et la réponse, toutes sortes d'hypothèses se présenteront à l'esprit de quelqu'un de plus expérimenté. Le médecin qui cherche à déterminer si son patient présente un cancer commencera par formuler des réponses contingentes. Il ne dira pas « oui, les données permettent de conclure » mais répondra : « si cette hypothèse est confirmée, alors je peux conclure qu'il a un cancer », si bien qu'il procédera à de nouvelles observations.

C'est pourquoi il est souhaitable que le Système Expert, au même titre que l'être humain, soit capable de produire des propositions contingentes. Face à une question difficile, dont il ne peut pas dire qu'elle soit vraie ou fausse en présence des données disponibles, il doit être en mesure d'exhiber les conditions qui rendront cette proposition vraie afin d'orienter les recherches futures. Ce problème consiste justement à identifier quels termes propositionnels doivent être vrais, ou de façon plus générales, quelles relations doivent exister entre ces termes propositionnels, pour que la formule mise à l'étude s'avère tantôt vraie, tantôt fausse.

Le problème qui consiste à trouver les conditions qui satisfont une formule normale conjonctive $C_1 \wedge C_2 \wedge \dots \wedge C_n$ se scinde en deux cas d'école. Dans le cas particulier où les clauses $C_1, C_2, \dots, C_n$ ne contiennent toutes qu'un ou deux termes, le problème s'appelle 2-SAT (où sat est une abréviation du mot « satisfaisabilité »). Dans le cas où une clause contient au plus trois termes, l'on parle du problème 3-SAT. Dans le cas général, où aucune restriction n'est posée sur le nombre de littéraux, on parle de CNF-SAT

(CNF étant un acronyme de *Conjunctive Normal Form* en anglais). Je vais décrire ces problèmes et la façon de les résoudre.

6.5.1 Complexité de 2-SAT

Le problème 2-SAT, relativement simple, peut être résolu en un temps linéaire. Ce que je veux dire par là, c'est que le nombre d'étapes de résolution nécessaires pour déterminer si 2-SAT est satisfaisable croît linéairement en fonction de la taille du problème. Par « taille du problème », j'entends le nombre total de clauses $C_1, C_2, \dots, C_n$ intervenant dans la formule à analyser. Par « étapes de résolution », j'entends n'importe quelle opération requise par un algorithme, telle que : lire un symbole, écrire un symbole, etc.

Formulé différemment, il existe une fonction linéaire entre le nombre d'étapes requises pour l'algorithme de résolution et n , le nombre de clauses comprises dans la formule $C_1 \wedge C_2 \wedge \dots \wedge C_n$. On dira, en notation mathématique, que le problème est de difficulté temporelle $\mathcal{O}(n)$, où la lettre stylisée $\mathcal{O}$ désigne n'importe quelle fonction linéaire prenant n comme argument.

Je vais donner ici un schéma de la preuve. Tout d'abord, il est important de remarquer que chaque clause $l_1 \vee l_2$ du problème 2-SAT peut se réécrire à l'aide d'une implication de deux manières différentes : $\neg l_1 \rightarrow l_2$ ou $\neg l_2 \rightarrow l_1$. Si on a par exemple la formule suivante :

$$(p_1 \vee p_2) \wedge (\neg p_2 \vee p_3) \wedge (\neg p_3 \vee p_1) \wedge (\neg p_1 \vee p_2)$$

À partir de ces quatre clauses, on peut écrire huit implications. Je les donne ci-dessous, après m'être débarrassé des éventuelles doubles négations ($\neg \neg p_i$ se simplifie p_i) :

- $\neg p_1 \rightarrow p_2$
- $\neg p_2 \rightarrow p_1$
- $p_2 \rightarrow p_3$
- $\neg p_3 \rightarrow \neg p_2$
- $p_3 \rightarrow p_1$
- $\neg p_1 \rightarrow \neg p_3$
- $p_1 \rightarrow \neg p_2$
- $\neg p_2 \rightarrow \neg p_1$

En associant les formules (1), (3) et (5), on peut dériver la chaîne d'implication suivante :

$$(\neg p_1 \rightarrow p_2) \wedge (p_2 \rightarrow p_3) \wedge (p_3 \rightarrow p_1)$$

Cette chaîne indique que si p_1 est faux, alors p_2 est vrai (première clause). Or si p_2 est vrai, p_3 doit être vrai (deuxième clause). Finalement si p_3 est vrai, on a p_1 vrai. Il s'ensuit que si p_1 est faux, alors p_1 doit être vrai.

On en conclut que p_1 doit être vrai. Il reste à se demander si le fait que p_1 soit vrai ne conduit pas, par un raisonnement similaire, à la conclusion que p_1 doit être faux. Or il se trouve qu'on peut dériver la chaîne suivante en associant les formules (7) et (8) :

$$(p_1 \rightarrow \neg p_2) \wedge (\neg p_2 \rightarrow \neg p_1)$$

On conclut que p_1 ne peut pas non plus être vrai. Il en découle que la formule est insatisfaisable.

Cette méthode illustre une stratégie pour résoudre le problème 2-SAT. Dans un premier temps, il faut écrire toutes les $2n$ implications associées aux n clauses. Ensuite, il faut vérifier l'existence d'une chaîne allant d'une variable vers sa négation et de sa négation vers la variable de départ.

L'approche naïve qui a été illustrée ci-dessus consiste, pour chaque variable p_i , à chercher d'abord s'il existe une chaîne conduisant de p_i à $\neg p_i$, puis de $\neg p_i$ à p_i . Pour trouver une chaîne allant d'une variable à une autre, on peut avoir jusqu'à $\mathcal{O}(n^2)$ étapes^[13]. Si on le fait pour chaque variable, la complexité monte à $\mathcal{O}(n^3)$. C'est pourquoi le nœud du problème consiste à structurer l'information de telle sorte qu'il ne soit pas nécessaire de parcourir la liste d'implications plus de k fois, où k est une constante ne dépendant pas de n .

Une solution a été suggérée en 1979 par Aspvall, Plass et Tarjan^[14]. Elle consiste à représenter les connections entre les termes propositionnels à l'aide d'un graphe dirigé. Ce graphe est une structure mathématique composée de points, aussi appelés nœuds, dont certains sont reliés par des flèches directionnelles. On s'assure que :

- Un point soit créé pour chaque terme propositionnel p_i figurant dans la formule de départ ;
- Un point soit créé pour la négation $\neg p_i$ de chaque terme propositionnel figurant dans la formule de départ ;
- Une flèche reliant les points x et y est créée si et seulement s'il existe une implication $x \rightarrow y$ (où x et y représentent chacun un terme propositionnel ou sa négation).

Ce graphe est représenté dans le système informatique en créant, pour chaque point x du graphe, une séquence de variables $\langle x, y_1, y_2, \dots, y_l \rangle$. La variable x représente le terme propositionnel associée à ce point du graphe. Les variables $y_1, y_2, \dots, y_l$ sont l'ensemble des points connectés à x au moyen d'une flèche sortant de x .

13. Une chaîne peut comporter, dans le pire des cas, près de $2n$ implications. Or chaque implication demande de parcourir la liste des $2n$ implications. La complexité totale est donc de l'ordre de $2n \times 2n$, soit $\mathcal{O}(n^2)$.

14. BENGT ASPVALL, MICHAEL F. PLASS ET ROBERT E. TARJAN, *A linear-time algorithm for testing the truth of certain quantified boolean formulas*, Information Processing Letters, vol. 8, no 3, 1979, p. 121-123.

Le problème consiste alors à créer ce graphe en $\mathcal{O}(n)$ opérations et à montrer s'il existe un cycle permettant d'aller d'une variable donnée à sa négation, puis de cette négation à la variable de départ. Ceci est possible en utilisant soit l'algorithme de Kosaraju, qui parcourt deux fois les points du graphe, ou l'algorithme de Tarjan, qui ne les parcourt qu'une seule fois. Je ne décrirai pas ces méthodes, mais elles consistent à organiser intelligemment l'information détenue sur chaque point.

6.5.2 Complexité de 3-SAT

Le problème 3-SAT présente des clauses contenant chacune, au plus, trois littéraux $l_1 \vee l_2 \vee l_3$. La méthode décrite dans le cas 2-SAT n'est plus valable car il n'est plus si évident de représenter cette clause uniquement au moyen d'implications. On devrait en effet commencer par écrire six implications :

- $\neg l_1 \rightarrow (l_2 \vee l_3)$
- $\neg(l_2 \vee l_3) \rightarrow l_1$
- $\neg l_2 \rightarrow (l_1 \vee l_3)$
- $\neg(l_1 \vee l_3) \rightarrow l_2$
- $\neg l_3 \rightarrow (l_1 \vee l_2)$
- $\neg(l_1 \vee l_2) \rightarrow l_3$

Chaque disjonction $l_2 \vee l_3$ intervenant dans ces équations doit être remplacée par $\neg l_2 \rightarrow l_3$ ou $\neg l_3 \rightarrow l_2$. Cela peut entraîner certaines difficultés. Tout d'abord, on peut assister à des expressions contenant la négation d'une implication, comme par exemple :

$$\neg(\neg l_2 \rightarrow l_3) \rightarrow l_1$$

Il n'est pas forcément évident de manipuler la négation d'une implication dans la représentation sous forme de graphe qui a été donnée plus haut. De plus, et il s'agit là d'une deuxième difficulté, les implications que nous manipulons sont des chaînes de trois littéraux, du type :

$$\neg l_1 \rightarrow (l_2 \rightarrow l_3)$$

La question de savoir s'il est possible d'adapter cette nouvelle forme du problème à l'algorithme de Tarjan ou Kosaraju, ou à toute autre méthode permettant une résolution de complexité $\mathcal{O}(n)$, ou même $\mathcal{O}(n^k)$, où k est un entier positif quelconque, est un problème ouvert doté par l'institut Clay d'une récompense d'un million de dollars. Les solutions connues actuellement sont du type $\mathcal{O}(2^n)$. Tout ce que l'on peut dire avec certitude à propos de 3-SAT, c'est que bien qu'il soit souvent difficile, en pratique, de trouver une solution (la méthode de résolution), il est en revanche assez évident de

vérifier si une réponse s donnée est une réponse satisfaisante au problème (méthode de vérification).

Cela revient à prouver qu'il existe une méthode de vérification polynomiale au problème. Si on donne en effet une certaine valeur aux termes propositionnels $p_1, p_2, \dots$ intervenant dans la formule, il suffit d'utiliser la table de vérité pour déterminer si la formule dans son ensemble est vraie ou fausse. Cette vérification demande un nombre d'étape proportionnelle au nombre de clauses dans la formule (rappelons que cette formule est écrite sous forme normale conjonctive) ; autrement dit, il suffit de balayer une fois la formule de gauche à droite. On en conclut que 3-SAT est donc de complexité $\mathcal{O}(n^k)$, où n est le nombre de symboles dans la formule et où k est un entier quelconque. Cela revient à dire que 3-SAT est de complexité polynomial.

D'un point de vue terminologique, tous les problèmes tels que 3-SAT, dont la méthode de vérification est de complexité temporelle linéaire, appartiennent à la classe de problèmes NP. L'acronyme NP vient de l'anglais Non-deterministic Polynomial et fait référence au fait qu'une machine de Turing non déterministe prendrait un temps polynomial pour le résoudre. La définition exacte entre machines déterministes et non déterministes sera abordée plus loin dans cette thèse. Qu'il suffise de savoir pour l'instant, et afin de rester bref, que la machine déterministe n'explore qu'un chemin à la fois tandis que la machine non-déterministe parcourt plusieurs chemins simultanément. Autrement dit, une machine déterministe peut essayer toutes les réponses et voir si elles fonctionnent, tandis qu'une machine déterministe doit justement trouver quelle est la réponse correcte.

Définition 10. *Un problème NP est un problème dont une solution donnée peut être vérifiée (rejetée ou acceptée) en un temps polynomial.*

6.5.3 Complexité de CNF-SAT

L'on pourrait se demander quelle est la complexité des problèmes 4-SAT, 5-SAT, ... k -SAT, où k est un entier positif quelconque ? Dans la mesure en effet où le passage de 2-SAT à 3-SAT changeait radicalement la nature du problème, il est permis de croire qu'il en soit de même en passant de 3-SAT à 4-SAT. Or il se trouve que ce n'est pas le cas. Il est possible de montrer que le problème k -SAT (avec $k > 2$) peut se réduire au problème 3-SAT, dans ce sens précis que si une méthode de résolution d'une complexité $\mathcal{O}$ donnée est détenue pour 3-SAT, cette même méthode fonctionnera pour 4-SAT sans changer l'ordre de complexité.

C'est cette propriété qui permet de réunir tous les problèmes de type 3-SAT ou supérieur sous l'intitulé CNF-SAT. Autrement dit, déterminer la satisfaisabilité d'une formule sous forme normale conjonctive, à l'exception du cas où ces clauses ne contiennent que 2 littéraux maximum par clause, c'est résoudre un seul et même problème général.

Le Théorème de dichotomie de Schaefer généralise ce résultat en montrant quelles sont les conditions précises pour qu'une variante du problème SAT soit assimilable, en termes de complexités, tantôt au problème 3-SAT, tantôt au problème 2-SAT.

6.5.4 Complexité spatiale

Outre la complexité temporelle dont nous venons de faire l'analyse, qui cherche à déterminer le nombre d'étapes nécessaires dans le pire des cas, il est important de garder à l'esprit qu'un algorithme utilise aussi des ressources de mémoire. Il garde pendant son exécution la trace d'un certain nombre de symboles. Si ce nombre est trop élevé, l'espace de stockage risque de déborder avant que l'algorithme ait convergé vers une solution. C'est pourquoi il est important de déterminer la relation existant entre la taille du problème et le nombre de symboles qui doivent être gardés en mémoire, à chaque instant du processus, dans le pire scénario.

On dit qu'un programme a une complexité spatiale $\mathcal{O}(n)$ si le nombre de symboles stockés en mémoire dans le pire des cas est proportionnel à la taille du problème. Ici, la taille du problème est donné par le nombre d'éléments propositionnels $p_1, p_2, \dots, p_n$. En 1970, Savitch a montré un théorème¹⁵ ayant pour implication que 2-SAT est de complexité spatiale $\mathcal{O}(\log(n))$. Cela signifie que le nombre de symboles qu'il faut mémoriser est une fonction logarithmique de n . Il s'agit de l'un des rares résultats permettant de dire qu'un problème de complexité temporelle déterminée possède une complexité spatiale. On ne connaît en effet pas la complexité spatiale de 3-SAT.

6.6 Problèmes NP-complets

À ce stade, j'aimerais anticiper une objection dans l'esprit du lecteur. Bien que j'ai montré que les problèmes k-SAT (avec $k > 2$) soient difficiles à résoudre, il reste à se demander s'ils constituent des problèmes d'un intérêt pratique. Il se pourrait en effet que la plupart des questions que nous posons à des Systèmes Experts puissent être représentés sous une forme normale conjonctive avec, au plus, deux littéraux par clause. Si tel était le cas, il n'y aurait rien d'embarrassant à ce que les problèmes 3-SAT soient plus complexes à résoudre que les problèmes 2-SAT. Or il se trouve que non seulement un grand nombre de problèmes intéressants sont de type 3-SAT, mais que ces problèmes forment une classe de problèmes réductibles les uns aux autres. Autrement dit, tous ces problèmes ont la même structure ; développer une méthode de résolution pour l'un d'entre eux, à un isomorphisme

15. Le théorème de Savitch, exprimé le plus simplement possible, stipule que si un problème est de complexité spatiale $\mathcal{O}(\log(n))$ avec une machine de Turing déterministe, il est aussi de complexité spatiale $\mathcal{O}(\log^2(n))$ pour une machine non déterministe.

près, c'est détenir une méthode pour chacun.

On doit à Stephen Cook et à Leonid Levin d'avoir démontré en 1971, de façon quasi simultanée, que toute méthode de résolution valide pour 3-SAT peut se transposer aux méthodes de la classe *NP*. Cette « transposition » elle-même se fait en temps polynomial, ce qui veut dire que la transposition n'est pas plus complexe à réaliser que ne l'est la méthode de résolution proprement dite (qui, jusqu'à preuve du contraire, est exponentielle). D'un point de vue formel, ce résultat est au fondement du concept de NP-complétude, dont le problème 3-SAT représente l'archétype par excellence. On dira qu'un problème est NP-complet lorsqu'il respecte les deux propriétés suivantes :

- **Il appartient à NP :** Le problème P doit être tel qu'il soit « facile » de vérifier si une réponse s donnée constitue une solution de ce problème P . Par « facile » j'entends que la vérification puisse être menée en un temps d'exécution polynomial. Autrement dit, il existe un algorithme qui, pour toute réponse possible s , détermine si oui ou non s est une solution du problème ; la complexité temporelle de cet algorithme est $\mathcal{O}(n^k)$, où k est en entier positif.
- **Il est transposable aux problèmes de NP :** Le problème P doit être tel que tout problème P' respectant la condition (1) puisse se réduire au problème P via une transformation polynômiale. Cela veut dire que si l'on détient une solution s pour le problème P , il existe une fonction f qui permet de trouver la solution s' du problème P' . Cette fonction doit être calculable en temps polynomial. Autrement dit, elle doit être de complexité $\mathcal{O}(n^k)$. En termes intuitifs, on dira du problème P qu'il est au moins aussi difficile que tous les problèmes de la classe *NP*.

Il est important de noter que les problèmes de la classe NP ne sont pas tous NP-complets. Certains problèmes appartenant à NP ne sont en effet pas réductibles aux autres problèmes de cette classe, ce qui veut dire que trouver un algorithme de résolution efficace pour ces problèmes particulier nous apprend rien sur les autres problèmes, comme 3-SAT.

Inversement, il existe possiblement des problèmes qui ne sont pas de la classe NP mais dont la méthode peut être transposée pour résoudre tous les problèmes de la classe NP. Un problème respectant cette propriété est dit NP-hard, au lieu de NP-complet. Les problèmes NP-hard doivent être tels que trouver une solution ou vérifier une solution sont, dans les deux cas, des opérations tout aussi difficiles. Jusqu'à présent, aucune preuve n'a été donnée que les problèmes NP-hard ne sont pas tous, en dernier ressort, des problèmes NP. En effet, le fait que nous ne connaissions pas d'algorithme

pour vérifier rapidement leurs solutions ne veut pas dire qu'il n'en existe pas¹⁶.

S'il y a tout de même bien une connaissance qui se fortifie avec le temps de façon certaine, cette liste de problèmes NP-complet. Depuis les années septante, un certain nombre de problèmes, jugés complexes, étaient connus en mathématiques discrètes, et représentés le plus souvent à l'aide de graphes, d'ensembles ou encore de systèmes d'équations linéaires : allocation de ressources, détermination d'un itinéraire optimal, remplissage d'un certain volume, etc. Richard Karp montre en 1972 que 21 de ces problèmes sont NP-complets et sont donc réductibles les uns aux autres. On en connaît aujourd'hui tellement qu'il serait difficile de les répertorier. Ils concernent tant la théorie des graphes que la théorie des jeux, la résolution d'équations, la géométrie euclidienne ou encore l'optimisation continue. Je vais en donner quelques exemples. Certains sont choisis au sein de la liste originale de Karp, tandis que d'autres sont plus récents :

- **Le problème du voyageur de commerce** : Un commerçant souhaite passer par un ensemble de villes données. Il connaît parfaitement la distance séparant chacune de ces villes les unes par rapport aux autres. Ces distances ne sont pas calculées à vol d'oiseau mais en fonction de la longueur des routes qui les relie effectivement. Il souhaite tracer un itinéraire tel que chaque ville soit visitée au moins une fois et que le nombre total de kilomètres parcourus soit inférieur à un entier k donné.
- **Le problème du sac à dos** : Un randonneur souhaite prendre un certain nombre d'objets dans son sac à dos, mais la taille de ce sac est limitée. Il assigne à chaque objet une valeur et un poids. Il cherche à maximiser la valeur totale des objets qu'il met dans le sac, sans que leur poids total ne dépasse la capacité du sac. Il existe une variante multidimensionnelle de ce problème, où chaque objet a un certain nombre d de paramètres (poids, largeur, longueur, hauteur...) et où le sac à lui-même une capacité différente pour chacun de ces paramètres.
- **Le problème de coloration de graphe** : Un data analyste souhaite attribuer une couleur ou un label à chaque membre d'un réseau social de sorte que deux personnes amies sur ce réseau n'aient pas la même couleur ou le même label. On peut représenter le réseau social en dessinant un point pour chaque personne inscrite et en reliant par

16. De manière générale, il est plus facile, en mathématique, de prouver l'existence d'un objet, que sa non-existence. Cette malédiction de la non-existence est responsable de ce que très peu de choses peut être affirmé avec certitude en ce qui concerne la complexité algorithmiques des problèmes combinatoires.

des segments de droite les points correspondant à des personnes qui sont amies entre elles. Dans le cas où aucun des segments de droite ainsi obtenus ne se croise, il suffit de quatre couleurs pour résoudre le problème. Ce résultat, connu sous le théorème des 4 couleurs, a une application pratique, connue de tous les écoliers : 4 couleurs suffisent pour colorier les pays d'une carte sans que deux pays limitrophes n'aient la même couleur.

- **Le problème de couverture par sommet** : Le roi Minos souhaite poster des gardes dans son labyrinthe de telle sorte que, avec un nombre minimum de personnes, il ait vue sur l'ensemble des couloirs. Les gardes ne doivent pas se déplacer. Ils restent dans une position fixe. En revanche, ils peuvent regarder dans toutes les directions. La version moderne de ce problème consiste à placer des caméras dans un bâtiment pour empêcher James Bond de s'y introduire à notre insu.
- **Le problème du cavalier** : Un joueur d'échecs place un cavalier sur un emplacement aléatoire du plateau de jeu. Il souhaite visiter les 64 cases en ne passant jamais deux fois au même endroit. Sous la plume de George Perec, ce problème peut aussi s'exprimer de la façon suivante : comment visiter les 100 emplacements d'un immeuble de dimensions 10 x 10 en ne passant jamais deux fois au même endroit ? La solution a donné naissance au livre « La vie, mode d'emploi », où chaque chapitre est consacré à une pièce de cet immeuble¹⁷. Mais si vous êtes plutôt féru d'architecture, vous préférerez sans doute la formulation d'Euler : est-il possible de flâner dans la ville de Königsberg, et de revenir à son point de départ, en passant par chaque pont exactement une fois ? Ce problème est légèrement différent, dans la mesure où il y a une contrainte en plus : celle de revenir au même endroit.
- **Le jeu du Démineur** : Ce célèbre jeu solo sur Windows, inventé initialement pour permettre aux utilisateurs de se familiariser avec l'usage de la souris, consiste en un plateau quadrillé. Chaque case peut soit contenir une bombe, soit un numéro correspondant au nombre de bombes présentes sur les cases adjacentes. Au début du jeu, les cases sont cachées. Le joueur peut les révéler une à une : s'il tombe sur une bombe, il perd la partie. Quiconque ayant déjà pratiqué ce « sport » se rend compte qu'il faut de temps en temps

17. Georges Perec commet en réalité une légère faute entre le 63ème et le 64ème déplacement, mais elle est probablement volontaire car la correction est aisée. Voir à ce propos <https://www.danthomasson.com/georges-perec.html>.

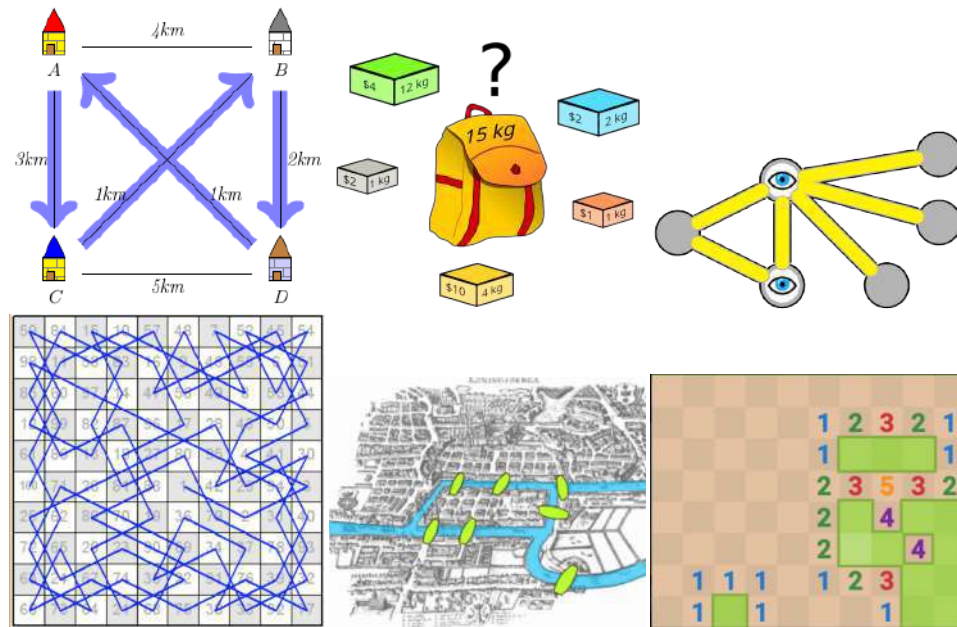


FIGURE 6.2 – De gauche à droite : (1) Le problème de voyageur de commerce, (2) Le problème du sac à dos, (3) le problème de la couverture par sommets, (4) Le problème du cavalier 10x10 de Georges Perec, (5) le problème des 7 ponts de Königsberg et (6) le jeu de Démineur.

« tenter sa chance ». Est-il cependant possible de remporter la partie sans s'appuyer sur le hasard ? Richard Kaye, de l'Université de Birmingham, a démontré en 2000 que toute configuration donnée du jeu définit un problème NP-complet^[18]. Cela veut dire que si vous voulez analyser le jeu pour éviter au maximum de « tenter votre chance », il va vous falloir résoudre un problème NP-complet. La preuve de cet intéressant résultat consiste à montrer qu'on peut construire n'importe quel problème 3-SAT à partir du jeu du Démineur^[19].

Il est facile de constater que ces problèmes touchent à de nombreuses applications pratiques. Or les résoudre avec un Système Expert revient à résoudre 3-SAT. Il n'existe actuellement aucun algorithme capable d'y ar-

18. KAYE R., *The Minesweeper is NP-complete*, The Mathematical Intelligencer, Volume 22, number 2, 2000, pp. 9–1.5

19. Pour être plus précis, Richard Kaye démontre qu'on peut utiliser le jeu du Démineur pour construire des portes logiques du type : « Négation », « Et », « XOR » (ou exclusif). En combinant ces portes, on peut construire une configuration du jeu du Démineur correspondant à une proposition logique sous forme normale conjonctive. Toute solution du Démineur est dès lors solution de ce problème 3-SAT.

river en temps polynomial, si bien que le calcul croît exponentiellement avec la taille du problème. C'est une des raisons, pas nécessairement la plus grave, pour laquelle le projet de réduire l'intelligence artificielle à la résolution d'équations logiques a rencontré de sérieuses difficultés à partir des années septante. Après une période d'engouement initial pour les Systèmes Experts, la question s'est posée de déterminer quelles méthodes alternatives pouvaient être utilisées pour faire face aux problèmes de la classe NP ?

Deux stratégies générales peuvent être envisagée. La première, et la plus simple à conceptualiser, consiste à abandonner les solutions exactes, au profit de solutions approchées. Dans certains cas, l'on peut se permettre d'assouplir les contraintes imposées sur la solution de sorte à étendre le champ de recherche. Cela dépendra bien sûr des applications, mais il semblerait que l'esprit humain est capable de supporter aisément l'inexactitude, de sorte à naviguer dans des contextes complexes, là où la machine refuse de procéder plus avant. Il se peut donc que la tolérance à l'inexactitude ou à l'ambiguïté soit une ressource précieuse à l'intelligence, entendue au sens le plus large.

La seconde approche, tout à fait complémentaire, repose sur une propriété essentielle des problèmes NP. On a dit en effet qu'il était toujours rapide (temps polynomial) de vérifier une solution de NP mais difficile (temps exponentiel dans les algorithmes actuels) de la trouver. Rien n'interdit dès lors un esprit intelligent de chercher à deviner la bonne réponse en recourant aux statistiques ou à son expérience. Il se trompera certes souvent, mais à force d'affiner son approche, et pour peu que le problème admette suffisamment de solutions, il n'est pas exclu qu'il finisse par converger vers quelque chose de robuste.

Cette stratégie adaptative, qui ne conçoit pas le problème en dehors de son contexte d'apprentissage, est également motivé par le fait que, quel que soit la méthode de résolution adoptée pour résoudre un problème NP, il faudra toujours partir d'un candidat initial et l'optimiser. Dans de nombreux cas de figure, le choix d'un bon candidat permettra à un algorithme, même médiocre, de converger en temps linéaire; alors qu'un mauvais candidat conduira le meilleur algorithme actuellement en vigueur à s'empêtrer dans des calculs de complexité exponentielle.

N'est-il dès lors par envisageable, en nourrissant la machine avec suffisamment d'exemples connus, de lui permettre d'améliorer progressivement sa capacité à "avoir de la chance" et à choisir le "bon" candidat initial ? Il semblerait que l'être humain ne procède pas différemment lorsqu'on le confronte à une tâche difficile. Prenons le problème du labyrinthe de Minos dont il était question plus haut. Ce problème consiste à placer un certain nombre de gardiens des corridors pour surveiller l'ensemble de la structure. Si Minos demandait à ses conseillers comment procéder, et que ceux-ci étaient des stratèges inexpérimentés, ils viendraient peut-être avec des réponses qui sont tout aussi mauvaises qu'en procédant au hasard. Mais si l'expérience était de leur côté, il est plausible de penser que les conseillers ne soumettraient

pas de trop mauvaises propositions à leur souverain. Une étude attentive permettrait ensuite de déterminer quelle est la meilleure d'entre elles — et de l'affiner autant que possible — mais on partirait, pour ainsi dire, avec un bon matériau de départ.

Cette capacité d'apprentissage que j'ai en vue, basée sur l'expérience, s'apparente peut-être au Daimôn de Socrate ? Le Daimôn en effet suggère à l'oreille du philosophe les voies impraticables. Il ne lui montre pas du doigt ce qu'il est bon de faire, mais il se réveille chaque fois que pied est posé sur un chemin aux conséquences morales douteuses. De la même manière, l'expérience ou l'intuition est une petite voix, en nous, qui oriente notre regard. Elle ne nous dit pas comment agir face à un problème, mais elle nous laisse avec un sentiment désagréable, un je-ne-sais-quoi d'incohérent, lorsque notre esprit se porte sur des approches douteuses. Au plus un homme acquiert de pratique, au plus il affinera cette voix en lui. Il ne soumettra plus à son jugement que les approches qui lui semblent dignes d'intérêt. Et même s'il arrive au stratège le plus expérimenté de louper la solution idéale, au moins ne perdra-t-il pas son temps à étudier les mauvaises.

Toutes ces considérations ont conduit l'intelligence artificielle sur une nouvelle voie. Au lieu de décrire aux algorithmes la marche à suivre de façon purement Top Down, il est devenu de plus en plus intéressant de leur donner la nourriture leur permettant de croître par eux-mêmes. Cette approche bottom-up, qui part des données brutes pour s'élever progressivement vers des calculs rationnels, est au cœur des réseaux de neurones. Le modèle d'une telle machine, ce n'est plus le juge impartial, qui décide sur base de lois logiques, mais c'est l'organisme vivant auto-adaptatif ; celui qui accumule des savoirs de façon dynamique. On appelle ce nouveau champ de recherche *l'intelligence Learning*. C'est de cela dont il va être question dans le prochain chapitre.

6.7 Critique épistémologique : Entre platonisme et logicisme

Mon ambition dans cette section était de montrer en quoi les stratégies top-down, incarnées le plus parfaitement par les systèmes experts, s'appuient sur le jugement analytique sans parvenir à capturer les formes *a priori* de la connaissance humaine. Lorsque Herbert Simon déclare que de tels systèmes seront capables, en moins de vingt ans, de réaliser n'importe quelle tâche cognitive à la portée d'un être humain²⁰, il fait implicitement l'hypothèse que l'intelligence humaine se réduit à un ensemble de déductions logiques, ou que la démarche hypothético-déductive puisse du moins reproduire, voire dépasser, les compétences de l'être humain.

20. Il écrit vers 1960 : « Machines will be capable, within 20 years, of doing any work that a man can do. » (The New Science of Management Decision)

Il est toutefois intéressant de noter que le jugement analytique dont il est question avec le General Problem Solver de Herbert Simon n'est pas celui qu'on avait repéré chez Platon, mais qu'il est au contraire plus proche de la raison calculante. C'est une machine spéculative détachée de toute sémantique. Ses opérations de pensée se réduisent à écrire des symboles sur le papier et à en dériver des conclusions rigoureuses.

Or Platon identifiait deux facultés rationnelles : la *dianoia* et la *noésis*. La première consistait à faire des inférences logiques à partir de certaines hypothèses, un peu à la façon des Éléments de géométrie d'Euclide. La *noésis*, quant à elle, c'était la contemplation de l'anhypothétique, ou autrement dit, de ce qui se trouve au-delà de toutes les hypothèses. Cet anhypothétique est souvent assimilé à l'idée du Beau, du Bien et du Vrai. Il se peut toutefois que ce ne soit qu'une analogie dans la pensée de Platon²¹.

Contempler une forme pure, pour Platon, c'est les contempler toutes. Dans le lieu intelligible (*topos noetos*), les Idées sont si bien mêlées les unes les autres qu'elles constituent un système cohérent. Ce lieu est l'hébertude du géomètre qui, d'un seul coup, est saisi par la beauté des figures mathématiques, l'harmonie de leurs relations, la symétrie, la concision, l'élégance sans nom des lois qui les gouvernent. La philosophie de Platon suit une dialectique ascendante et descendante. Ascendante d'abord, car elle part des formes sensibles et cherche à savoir de quelles idées elles participent — en observant des dessins de triangles dans le sable, il s'agit de saisir l'idée en soi du triangle. Descendante également, car elle cherche aussi, en partant de l'Idée la plus générale, à établir des distinctions d'espèce entre les concepts.

Les systèmes experts semblent avoir pris au sérieux ce double mouvement, mais en faisant l'impasse sur cette *noésis*. Ils possèdent certes des sortes de principes fondateurs : ce sont les axiomes de la théorie et les lois logiques. Il est même permis de chercher de nouvelles vérités sur base de ces principes, soit de façon ascendante, soit de façon descendante. L'usage de ces mots « ascendant » et « descendant » est cependant l'exact inverse de ce que Platon avait vu en eux. L'approche ascendante des systèmes experts consiste en effet à partir des principes premiers pour déduire des propositions vraies, en créant des distinctions. L'approche descendante, quant à elle, consiste à partir des propositions et à les prouver en les ramenant aux axiomes. Dans certains cas aussi, cela permet d'ajouter des clauses à la théorie.

Tout se passe donc comme si le système expert devait déjà détenir les principes premiers pour « penser ». Au lieu d'observer le monde sensible pour en tirer des vérités intelligibles, il évolue dans un monde sans contact

21. Il y a un débat opposant deux interprétations. Selon l'interprétation ontologique, l'Idée du Bien est réellement l'anhypothétique. Selon l'interprétation épistémologique, ce n'est que par analogie que Platon le compare à l'Idée du Bien, tout comme il compare l'Idée du Bien au soleil. Voir à ce titre ROUX SYLVAIN, *Chapitre III. Le bien comme principe dans la République*, in *La recherche du Principe chez Platon, Aristote et Plotin*, Vrin, 2004, p. 69-116, <https://books.openedition.org/vrin/5432?lang=fr>

avec l'expérience. La machine peut certes recevoir des données pour en tirer des conclusions, mais ces données doivent être formatées de façon logique. Il n'est pas possible de demander à la machine d'analyser une photo et de créer ses propres catégories, comme c'est le cas aujourd'hui avec les réseaux de neurones. Il n'y a du reste pas de capacité d'autoapprentissage : tant que vous ne modifiez pas les règles, le système reste, de toute éternité, égal à lui-même.

En compensation, le programmeur contrôle parfaitement la conduite de la machine. Ses conclusions sont traçables et justifiables : il est possible de prédire comment elle va se comporter dans de nombreux contextes et de comprendre quel principe, quelle clause, était responsable de ses agissements. Il n'y a cependant pas de contemplation de la vérité, au sens d'un exercice visant à transformer son regard sur le monde. La machine ne s'éduque vraisemblablement pas au sens moral et intellectuel, pour s'élever peu à peu dans la compréhension d'un idéal, d'une vérité intelligible.

On peut dire, pour résumer, que l'oubli de l'Entendement, amorcé avec la crise des mathématiques, s'est accentué dans les premiers programmes de l'intelligence artificielle, pour faire de la Raison, non plus la faculté de juger à propos des grandes idées, mais une aptitude à calculer. Cette version réduite de la raison platonicienne est ce que j'appellerai la raison numérique, ou encore : la rationalité moderne. Ce que la raison perd du côté de l'épistémologie, elle le gagne en normativité. Qu'est-ce que je veux dire par là ? Que le programmeur du système expert peut décider entièrement du vrai du faux. Il encode les lois de la pensée et les axiomes de la théorie ; il vérifie en outre les réponses. Les systèmes experts sont traçables (on peut vérifier les étapes de raisonnement) et vérifiables (on peut dire si les réponses sont vraies ou fausses).

En dépit de ces avantages évidents, il convient de se demander comment ont évolué les paradigmes de programmation et s'ils sont parvenus à intégrer le monde sensible dans leurs calculs. Comme disait Kant en effet, la raison ne doit pas tourner à vide. Elle a besoin de s'appuyer sur des objets de l'espace et du temps. Elle doit partir de l'expérience et se hisser vers les concepts de façon bottom-up.

La première stratégie de programmation à avoir vu le jour est désignée sous le nom de stratégie *top-down*. Elle consiste à déterminer des règles de raisonnement, à les encoder sous forme d'un algorithme et à les exécuter avec une machine. Cette approche peut être vue comme une procédure de résolution d'équations logiques, écrites sous formes de clauses "A et B", "A ou B", etc.

Cette procédure reçoit le nom de système expert, mais son expertise se limite aux règles qui lui ont été enseignées. Le système expert présente en effet plusieurs lacunes :

1. Il ne possède pas d'autonomie ou d'autocritique.
2. Il se montre assez insatisfaisante face à certains problèmes coûteux en ressources de calcul, notamment les problèmes NP-complets.
3. Il réduit la rationalité à la simple manipulation de symboles. Il fait l'impasse sur les dimensions esthétiques et universaliste de la raison humaine, véhiculant une vision de la rationalité fondée sur le calcul.

Dans le prochain chapitre, je me demanderai s'il existe une autre stratégie de programmation.

Chapitre 7

Approche Bottom-Up : Les Réseaux de neurones

*Les émotions sont imbriquées
dans les réseaux neuronaux de la
raison.*

Antonio Damasio

Les systèmes experts ont, pendant longtemps, constitué l'archétype de l'intelligence artificielle. Ces logiciels sont les Sherlock Holmes de la connaissance humaine : à partir d'un petit nombre de principes fondamentaux, ils sont capables de déduire, d'inférer, d'exclure des possibilités qui n'auraient pas forcément pu être envisagées par des esprits humains. Les règles de raisonnement et les principes utilisés par ces machines doivent néanmoins être encodés par un programmeur. Ces règles doivent pouvoir s'énoncer de façon précise et opérationnelle. Elles doivent également faire consensus de la part des personnes qualifiées dans la discipline.

Or dès 1943, alors que les systèmes experts n'étaient pas encore nés, MacCulloch et Pitts ont proposé une autre façon d'envisager l'intelligence artificielle. Dans un article intitulé « A logical calculus of the ideas immanent in nervous activity », ils partent de l'hypothèse que les capacités d'apprentissage humaines sont liées à une certaine façon dont les neurones reçoivent, traitent et transmettent de l'information. À partir de là, ils tentent de comprendre et de modéliser les propriétés mathématiques du neurone à partir des données connues sur le fonctionnement physiologique du cerveau^[1] :

En raison du caractère « tout ou rien » de l'activité du neurone, les événements neuronaux et les relations entre eux peuvent être traités au moyen de la logique propositionnelle. Il se trouve que le

1. McCULLOCH, W.S., PITTS, W., *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics 5, 115–133 (1943).

comportement de chaque réseau peut être décrit dans ces termes, avec l'adjonction de procédés logiques plus sophistiqués en présence de réseaux contenant des cycles.

En dépit d'une simplification assez marquée des processus biologiques en œuvre dans le fonctionnement du cerveau, et peut-être aussi en raison de celle-ci, le modèle de McCulloch et Pitts est un point de départ incontournable dans l'élaboration d'algorithmiques bottom-up. De nombreuses variantes ont été introduites depuis lors, par des auteurs tels que Hopfield (prix Nobel 2024), Kohonen ou Héroult-Jutten. Il est toutefois admis aujourd'hui que les modèles mathématiques du réseau de neurones, quoiqu'ils soient commodes pour générer un apprentissage informatique, ne représentent pas une théorie valide des processus neurologiques.

Dans ce chapitre, je vais commencer par décrire le fonctionnement du neurone biologique, après quoi j'introduirai le modèle mathématique de McCulloch et Pitts. J'exposerai ensuite différentes améliorations du modèle initial, spécialisées dans la perception d'images ou dans la rétention d'informations. Je terminerai en illustrant au moyen du logiciel AlphaGo les différences fondamentales existant entre les réseaux de neurones et les Systèmes Experts. Cela permettra, dans le prochain chapitre, de tirer quelques conclusions épistémologiques.

7.1 Le Neurone biologique vs formel

L'histologie dominante du *XIX^{ème}* se représentait le corps cérébral comme un magma de cellules fusionnées. Depuis les travaux de Santiago Ramon y Cajal, lauréat du prix Nobel en 1906, il est admis que le neurone est une unité fonctionnelle autonome. Je vais me contenter de donner ici quelques caractéristiques générales de son fonctionnement. Je me concentrerai principalement sur les aspects du neurone qui sont les plus pertinents pour comprendre les liens existant entre la neurophysiologie et l'informatique théorique. Je laisserai donc de côté les terminologies concernant les différents types de cellules (cellules de Purkinje, astrocytes, épendymocytes...) ou le lien exact que le système nerveux entretient avec le reste de l'organisme (mécanisme de nociception, rôle de la moelle épinière ou des fameux neurones de l'estomac...).

L'analogie souvent utilisée pour décrire le fonctionnement des neurones à sa grand-mère consiste à se représenter une masse de câbles entremêlés les uns aux autres (il faut cependant admettre, pour le bienfait de cette métaphore, que votre grand-mère n'est pas électrophobe). Chaque câble possède une extrémité réceptrice et une extrémité émettrice. Les extrémités réceptrices sont connectées aux extrémités émettrices. Il existe cependant un espace entre les câbles, de sorte que le courant ne peut pas circuler tant que le contact n'a pas été établi. Ceci peut être fait lorsqu'une goutte d'eau vient combler le vide les séparant. Ces gouttes sont secrétées par l'extrémité

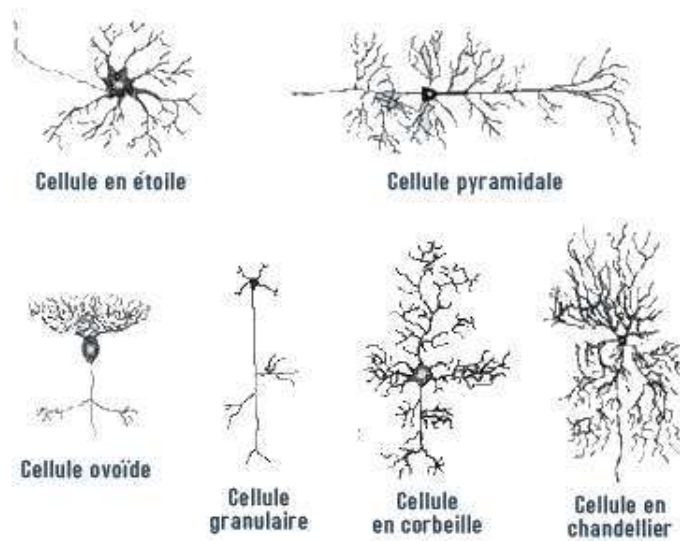


FIGURE 7.1 – Les neurones peuvent prendre différentes formes.

des câbles eux-mêmes en présence d'un potentiel électrique. Certaines de ces gouttes ont des formes qui leur permettent d'établir plus efficacement le contact entre certains câbles plutôt que d'autres. Lorsque deux câbles entrent en contact par l'entremise d'une goutte d'eau, une décharge électrique est susceptible de se produire. Un signal électrique se transmet ainsi d'un câble à l'autre.

Dans cette métaphore, les câbles représentent chacun un *neurone*, l'espace entre les câbles sont les *fentes synaptiques* et les gouttes d'eau sont les *neurotransmetteurs*. Les neurones sont des cellules nerveuses composées d'un noyau, d'un cytoplasme et d'une membrane. Le noyau possède le bagage génétique, le cytoplasme est riche en protéines et la membrane assure la cohérence de la cellule. Le neurone, et il s'agit là de sa spécificité, possède en outre des terminaisons tentaculaires, qui lui donnent tantôt la forme d'une étoile, d'une pyramide, d'une corbeille ou d'un chandelier. Ces terminaisons sont de deux espèces. Celles qui reçoivent des signaux électrochimiques en provenance d'autres cellules s'appellent des dendrites. Les prolongements qui diffusent le signal du neurone vers d'autres cellules sont appelés axones. Il y a donc bien une extrémité réceptrice et une extrémité émettrice (voir figure 7.2).

Les neurotransmetteurs, représentés par une goutte d'eau dans notre analogie, sont des molécules synthétisées par les neurones, ainsi que par d'autres cellules du système nerveux, comme les cellules gliales. Il existe une centaine de neurotransmetteurs connus, répartis en différentes familles : les acides aminés (glycine, glutamate...), les monoamines (dopamine, adrénaline, sérotonine...) et les peptines (opioïdes...), pour ne citer que les plus

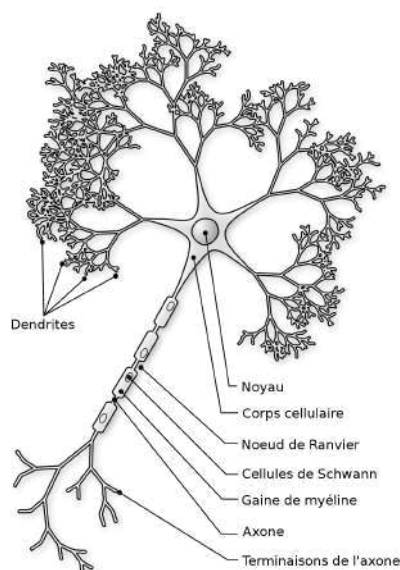


FIGURE 7.2 – Schéma biologique du neurone.

importantes. Il est courant d'associer à ces neurotransmetteurs un rôle physiologique et psychologique précis. La dopamine, par exemple, est décrite comme responsable du plaisir ; l'adrénaline contribue à la gestion du stress ; la sérotonine est l'hormone du bonheur, etc.

Il convient toutefois de se méfier de ces simplifications — que d'aucuns désignent sous le nom de neuromythes. S'il est avéré que les traitements de la dépression, par exemple, consistent à augmenter par voie chimique la concentration en sérotonine, il est scientifiquement dangereux de réduire chaque neurotransmetteur à un rôle unique. Lorsque l'on associe des concepts proprement humains (le bonheur, le plaisir, l'énergie...) à des unités physiologiques (sérotonine, dopamine, adrénaline...), on suppose qu'il existe une équivalence directe entre nos variables psychologiques et des cellules observables. Or il n'y a pas de raison de penser que la carte que nous utilisons, nous autres êtres humains, pour naviguer au travers de notre existence psychosociale, soit similaire à la carte que la nature utilise pour assigner des fonctions aux différentes cellules de notre organisme.

En guise d'illustration, une carence en adrénaline dans la substance noire, un noyau situé dans le tronc cérébral, peut provoquer des symptômes parkinsoniens (tremblement des mains) ; tandis qu'un excès de concentration de cette même adrénaline dans la voie mésolimbique peut mener à des symptômes schizophréniques (crises psychotiques). Ce double effet de l'adrénaline est loin de la représentation caricaturale qu'on lui impute généralement, lorsque l'on dit « aimer l'adrénaline » ou que l'on parle d'une « poussée d'adrénaline ». Le phénomène est d'ailleurs bien connu des pharmacolo-

gistes : lorsqu'ils prescrivent un médicament psychoactif, ils prennent le soin d'éviter de tomber de Charybde en Sylla, en provoquant des réactions schizophréniques chez un patient victime de Parkinson.

En gardant à l'esprit ces quelques réserves, que peut-on dire du rôle joué par les neurotransmetteurs dans l'activation neuronale ? Les neurotransmetteurs opèrent, grosso modo, comme des catalyseurs. Au plus ils sont présents dans une fente synaptique, au plus grandes sont les chances de laisser passer le courant d'un neurone vers un autre. Lorsqu'un neurone A est excité, il libère donc des neurotransmetteurs dans les fentes synaptiques des neurones $B_1, \dots, B_n$ auxquels il est connecté par des terminaisons dendritiques. Cette émission de neurotransmetteurs est généralement insuffisante, à elle seule, pour provoquer l'excitation de ces neurones $B_1, \dots, B_n$. Mais lorsque plusieurs neurones incidents à un neurone B s'excitent simultanément, B va s'exciter à son tour. De proche en proche, l'excitation neuronale va donc se propager au travers du cerveau.

Autrement dit, et il s'agit là d'une loi fondamentale, un neurone ne va s'activer que s'il est mis en contact avec suffisamment d'autres neurones excités. Cette règle, très importante pour ce qui va suivre, signifie que des chemins électriques variables peuvent se créer d'un neurone à l'autre. L'apprentissage du cerveau consiste à associer certains chemins neuroélectriques à certains contextes spécifiques, de telle sorte que chaque fois qu'une situation similaire se produit, un chemin similaire se crée. C'est ce phénomène qui explique que nous mémorisons plus aisément des informations liées à des informations déjà connues, plutôt que des informations totalement nouvelles. Cette idée de l'apprentissage par association statistique des activations des neurones est, du reste, au fondement du concept mathématique du « neurone formel ».

7.1.1 Le Neurone formel

Le neurone de McCulloch et Pitts est un modèle simplifié du fonctionnement du neurone biologique, qui fait l'économie de toute la complexité des processus physiologiques, pour ne conserver que quatre grandes idées :

1. Le neurone est une unité fonctionnelle qui reçoit le signal d'un ou plusieurs neurones (appelés neurones incidents) et qui envoie un signal à un ou plusieurs neurones (appelés neurones sortants) ;
2. Le signal envoyé est représenté sous forme d'un nombre binaire ou d'un nombre réel ;
3. Lorsque la somme des signaux reçus par un neurone, le potentiel p dépasse un certain seuil θ , ce neurone s'active et envoie un signal aux neurones sortants auxquels il est connecté ;
4. Les signaux émis par les neurones incidents sont multipliés par un coefficient synaptique, qui modélise l'importance de la liaison entre les

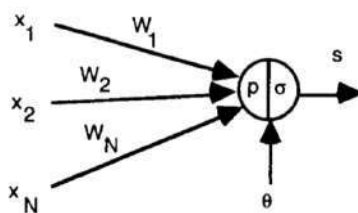


FIGURE 7.3 – Le neurone formel de McCulloch et Pitts (1943).

deux neurones.

D'un point de vue mathématique, le neurone reçoit les entrées $x_1, x_2, \dots, x_n$ et calcule le potentiel p . Celui-ci est donné par la moyenne pondérée des entrées : $p = x_1 W_1 + x_2 W_2 + \dots + x_n W_n$, où $W_1, W_2, \dots, W_n$ sont les « poids synaptiques » propres au neurone. Ces paramètres correspondent, dans le modèle biologique, à la réceptivité des dendrites. Plus le poids W_i est élevé, plus le signal x_i auquel ce paramètre est associé sera important dans la réponse du neurone. Si au contraire W_i est égal à zéro, alors le signal x_i n'est pas pris en compte : tout se passe virtuellement comme s'il n'y avait pas de connexion entre le neurone et l'entrée en question.

La réponse d'un neurone formel est binaire : soit le neurone s'active, soit il ne s'active pas. En termes mathématiques, cela signifie que le signal sortant du neurone, noté y , est une fonction définie pour toutes les valeurs réelles et ayant pour image les réponses -1 (inactivité) ou 1 (excitation). La valeur -1 est observée si le potentiel d'action p est inférieur à un seuil θ . La valeur 1 survient lorsque le potentiel d'action est inférieur au seuil θ :

$$y = \begin{cases} 0 & \text{si } p < \theta \\ 1 & \text{si } p \geq \theta. \end{cases}$$

Il est à noter que certains neurones formels ne renvoient pas la réponse -1 ou 1 , mais les réponses 0 ou 1 . Ce choix va dépendre des applications rencontrées et de certaines conventions concernant les variables utilisées. Il est également permis d'utiliser une fonction de décision pour produire des réponses réelles situées entre -1 et 1 . Cette approche est particulièrement intéressante si la réponse attendue n'est pas de type tout ou rien. C'est également une bonne méthode si le neurone A produit un signal qui va être lu par un neurone B. Ce signal sortant de A peut être plus ou moins important, ce qui va activer plus ou moins sensiblement le neurone B.

Le modèle de McCulloch et Pitts peut donc prendre des formes différentes selon le nombre d'entrées $x_1, x_2, \dots, x_n$, la valeur des poids $W_1, W_2, \dots, W_n$ et la fonction de décision liant la somme pondérée des entrées à la sortie y . Le nombre d'entrées dépend essentiellement des problèmes abordés : en reconnaissance d'images, il y aura trois entrées par pixel, une pour chaque

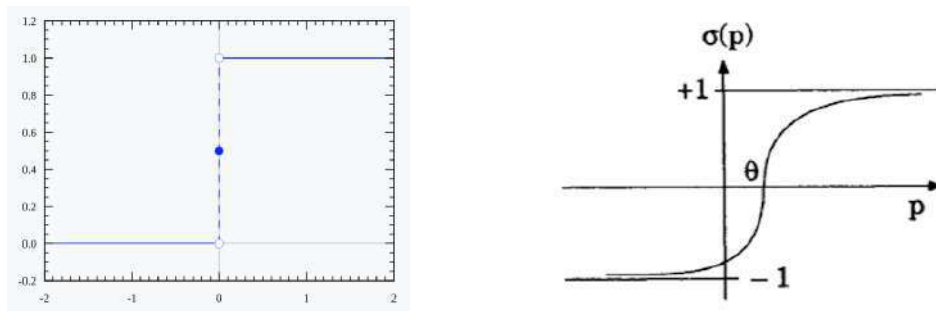


FIGURE 7.4 – La fonction de Heaviside, à gauche, ne peut prendre pour valeur que -1 ou 1 . La fonction sigmoïde, au contraire, produit des sorties continues entre ces deux points extrêmes. Sur le graphique à droite, la fonction de décision est $2\text{sig}(x) - 1$ de sorte à avoir des valeurs comprises entre -1 et $+1$

couleur fondamentale ; en classification de données, on aura une entrée pour chaque donnée, etc.

La fonction de décision, quant à elle, va être construite en regard du problème que l'on cherche à résoudre. Si la sortie du neurone doit délivrer une valeur binaire, correspondant par exemple à l'ouverture ou à la fermeture d'une porte, il faut utiliser une fonction de Heaviside, qui produit uniquement la réponse -1 ou $+1$ (voir figure 7.4) ; mais si la sortie est une valeur réelle, comme pour déterminer la température ambiante à l'aide de capteurs, on peut opter pour une fonction sigmoïde, dont la caractéristique entrée/sortie est, au demeurant, plus proche de celle du neurone réel (voir ??).

Dans un modèle de réseaux de neurones formels, l'apprentissage consiste à calculer les valeurs des coefficients synaptiques $W_1, W_2, \dots, W_n$ en se basant sur les exemples disponibles. À l'opposé des méthodes top-down, il n'est pas nécessaire d'isoler heuristiquement les grands principes fondamentaux de résolution du problème. On procède de façon aveugle, en donnant à la machine les paramètres formels d'apprentissage. Ces paramètres sont construits de façon à ce que les réponses du modèle coïncident avec les solutions connues du problème. Au plus ces données sont précises et nombreuses, en formant un échantillon statistiquement varié, au plus le modèle résultant sera performant. Il n'est donc pas rare que la machine résultant de cette approche adopte des comportements parfois difficiles à déchiffrer pour un être humain. L'on parle alors d'opacité intrinsèque.

En guise d'illustration, on pourrait imaginer un neurone formel qui aurait pour tâche de décider si l'âne de Buridan doit manger du foin ou de l'avoine. Ce neurone prend deux entrées x_1 et x_2 , quantifiant respectivement son appétit et à son enrhumement. Plus x_1 est élevé et plus l'âne a faim. Au plus x_2 est élevé et au plus l'âne est sensible aux crises de toux. Ces deux

valeurs x_1 et x_2 ne connaissent pas de limite positive, mais doivent être égales ou supérieures à zéro.

La réponse de l'âne — à son plus grand dam — est une réponse binaire : soit $y = 1$, dans quel cas l'âne mange du foin ; soit $y = -1$, dans quel cas il mange de l'avoine. On peut donc prendre pour fonction de décision la fonction suivante :

$$y = \begin{cases} -1 & \text{si } p = W_1x_1 + W_2x_2 < 0 \\ 1 & \text{sinon.} \end{cases}$$

Cet exemple correspond au perceptron imaginé par Frank Rosenblatt en 1957, qui a pour but de séparer un ensemble de données en deux classes disjointes. Dans le cas présent, ces classes correspondent à « manger du foin » ou « manger de l'avoine ». Les poids synaptiques W_1 et W_2 doivent être choisis de sorte à prédire la réponse de l'âne. On pourrait supposer que le foin est plus nourrissant que l'avoine, si bien qu'au plus grand est l'appétit de l'âne, au plus il se tournera vers le foin. De même, on pourrait estimer que l'avoine provoque moins d'éternuements que le foin, de sorte qu'au plus l'âne est enrhumé, au plus il se tournera vers l'avoine. Une solution simple consisterait alors à prendre W_1 positive et W_2 négative. Il existe cependant beaucoup d'autres choix possibles. Lesquels sont les plus judicieux ? D'après quels critères ?

7.2 La Règle de Hebb : qui se ressemble s'assemble !

J'ai donné la définition du neurone formel de McCullocht et Pitts et je l'ai comparée au neurone biologique. J'ai montré que tout l'enjeu des neurones artificiels est de créer des associations pertinentes entre les entrées $x_1, \dots, x_n$ et la sortie y . La question qui s'impose, à ce stade, consiste à comprendre comment exactement les poids synaptiques W_1 et W_2 doivent être fixés pour permettre au neurone de réaliser certaines tâches d'apprentissage ? Dans l'exemple de l'âne de Buridan, on se demande comment le neurone formel de notre âne peut lui permettre de faire de meilleurs choix. Pour y parvenir, nous aimerions utiliser les données qui sont déjà à notre disposition : à savoir, toutes les expériences passées dans lesquelles l'âne, après avoir senti un certain appétit et un certain degré d'enrhument, s'est tourné vers l'une ou l'autre substance comestible. Nous aimerions répéter les choix qui se sont avérés judicieux et inhiber les autres.

Dans son ouvrage intitulé *L'organisation du comportement*, publié en 1949, le docteur Hebb formule une règle à propos du comportement des neurones qui s'est avérée d'un grand retentissement pour résoudre le problème qui nous intéresse. Cette règle, aujourd'hui connue sous le nom de « règle de Hebb », stipule que lorsque deux neurones s'excitent ensemble, ils restent liés l'un à l'autre :

Quand l'axone d'une cellule A est assez proche pour exciter une cellule B et qu'elle prend part de façon répétée ou permanente à son activation, un processus de croissance ou un changement métabolique prend place dans l'une de ces cellules, ou les deux, de sorte que l'efficacité de A, en tant que cellule excitatrice de B, est augmentée.

Cette loi physiologique, si nous la transposons au modèle de McCulloch et Pitts, suggère d'accroître la valeur des coefficients synaptiques uniquement entre les neurones formels qui ont une activité synchronisée, et uniquement ceux-là. Pour comprendre comment cela peut être fait dans le détail, je souhaiterais illustrer la méthode avec un seul neurone formel, puis l'étendre à un réseau complet.

7.2.1 Neurone isolé

Commençons par envisager que notre âne de Buridan, tout juste sorti des mains de Dieu, soit doté d'un seul et unique neurone. Ce neurone est connecté à deux entrées : l'appétit x_1 et l'enrhumement x_2 . Il produit une réponse alimentaire y qui vaut 1 (manger du foin) ou -1 (manger de l'avoine). Conformément à la Règle de Hebb, le neurone va tenter de renforcer les habitudes existantes. Si notre animal mange souvent de l'avoine lorsqu'il est enrhumé, le neurone doit donc privilégier la réponse $y = -1$ lorsque x_2 est élevé.

D'un point de vue mathématique, on peut représenter toutes les expériences passées de notre âne, ses préférences antérieures, à l'aide de trois points x_1, x_2 et y . Les entrées x_1 et x_2 peuvent être placées dans un plan à deux dimensions (voir figure 7.5). Ces points (x_1, x_2) sont de couleur rouge lorsqu'ils sont associés à la réponse $y = 1$ et de couleur bleue lorsqu'ils sont associés à la réponse $y = -1$. Sous cette représentation, le rôle du neurone consiste à tracer une frontière entre les points rouge et bleu au moyen d'une droite de régression statistique, indiquée en pointillés.

Après avoir tracé la frontière, on dit que le neurone a terminé sa phase d'apprentissage ; il est prêt à prédire les comportements futurs de l'âne. Lorsqu'il recevra une entrée (x_1, x_2) inconnue, il répondra 1 si celle-ci est située au-dessus de la droite et -1 si elle est en-dessous de cette droite. Autrement dit, le neurone utilise les expériences passées pour prédire les comportements futurs.

Pour obtenir cette fameuse droite de régression qui tient lieu de frontière entre $y = 1$ et $y = -1$, plusieurs méthodes existent. La plus connue, qui vous rappellera sans doute vos cours de statistiques, consiste à chercher la droite D qui minimise la somme des carrés des distances. Intuitivement, D est la droite qui passe à l'exact milieu entre l'ensemble des points bleus et l'ensemble des points rouges. On dira des points utilisés pour l'apprentissage du neurone qu'ils appartiennent à sa base d'apprentissage. Les points

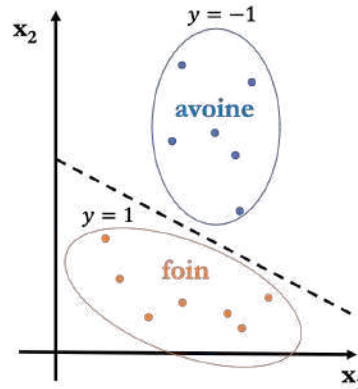


FIGURE 7.5

nouveaux, en revanche, sont des prédictions.

La droite de régression statistique est une séparation généralement efficace, mais pour en calculer la pente, il faut avoir sous les yeux l'ensemble des points rouges et bleus de la base d'apprentissage. Cela veut dire que, chaque fois qu'une nouvelle donnée est ajoutée à cette base, il faut recalculer la pente de la droite dans son entier. Or le temps de calcul croît avec le nombre de points. En présence de big data, contenant parfois des millions de points, notre neurone se montrera peu apte à assimiler de nouvelles informations. Il n'est donc pas assez flexible pour s'adapter aux contextes nouveaux, comme on le voudrait au sein de l'approche bottom-up.

Pour remédier à cette difficulté, le modèle de l'ADALINE, créé en 1960 par Bernard Widrow et Ted Hoff, prend une toute autre approche. Plutôt que de regarder d'un seul coup l'ensemble des points de la base d'apprentissage, il les considère un par un. Il utilise les premiers points pour tracer une ligne de séparation très approximative, forcément erronée. Chaque fois qu'un nouveau point est envisagé, deux scénarios sont possibles. Soit ce point se trouve du bon côté de la droite, dans quel cas il ne se passe rien. Soit le point est du mauvais côté de la droite, dans quel cas le neurone change légèrement l'orientation de cette droite de manière à minimiser l'erreur. De proche en proche, la droite va progressivement converger vers la méthode des moindres carrés. L'ADALINE a donc la même efficacité qu'une régression statistique standard, mais avec un apprentissage dynamique.

Si l'apprentissage est bon et que les données présentent une bonne structure, le neurone est capable de retrouver les réponses $y(i)$ associées aux points $x(i)$ de sa base d'apprentissage. Il y a donc bien association automatique des entrées apprises avec les sorties correspondantes. Ce qui nous intéresse cependant au premier chef, c'est ce que le neurone va effectuer lorsque l'entrée n'est pas exactement égale à $x(i)$, mais qu'elle s'en rapproche ? La plupart des situations de la vie ne sont en effet pas des copier-coller exacts

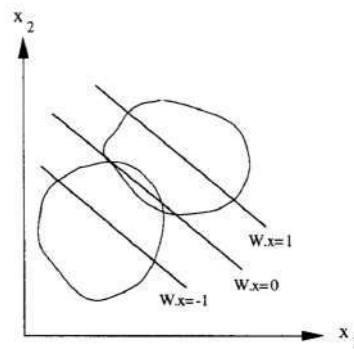


FIGURE 7.6 – Exemple de classes non linéairement séparables.

du répertoire d'apprentissage, sans quoi on ne pourrait pas parler d'adaptation intelligente de la part du neurone, pas plus qu'on ne dirait d'un joueur d'échecs qu'il est intelligent s'il se contente de jouer de mémoire les coups appris dans ses ouvertures. Pour le dire encore autrement, tout l'intérêt de l'intelligence artificielle bottom-up consiste en ceci que le neurone va « improviser » face aux situations nouvelles, en essayant de reconduire les inconnues au plus près de ce qu'il connaît déjà.

Comme les équations d'association que nous avons établies sont linéaires, les réponses du neurone seront des combinaisons linéaires des données mémorisées. En termes techniques, on dit du neurone qu'il procède à une régression statistique du premier degré. Bien que cela représente une bonne approximation, cette approche peut laisser à désirer sur plusieurs aspects. Tout d'abord, il existe des classes de données qui ne sont pas linéairement séparables, ce qui signifie qu'aucune méthode linéaire ne peut déterminer sans erreur si une entrée x doit être associée à la sortie s_1 ou à la sortie s_2 (voir graphique). Pour le dire en des termes imagés — et pour reprendre les termes de l'exemple précédent — cela signifie qu'il existe des situations où l'âne de Buridan ne peut pas décider, sans risque d'erreur, si le foin est meilleur que l'avoine (voir figure 7.6).

Un autre problème des méthodes linéaires vient de ce qu'elles sont moins robustes face aux changements de contexte. Idéalement, on voudrait que le neurone formel soit résilient : lorsqu'il est exposé à des couples de points (x_1, x_2) auxquels il n'est pas habitué, on souhaite que sa prédiction reste robuste. Un manque de résilience se manifeste, par exemple, lorsqu'un petit changement de paramètres de luminosité invalide un algorithme dédié à la reconnaissance d'image, qui échoue à reconnaître la tour Eiffel dans l'obscurité (voir figure 7.7).

En des termes techniques, si le vecteur x passé en entrée est proche d'un vecteur $x(i)$ qu'il a mémorisé, le neurone devrait produire une sortie y identique ou proche à la sortie mémorisée $y(i)$. Dans le cas de régression

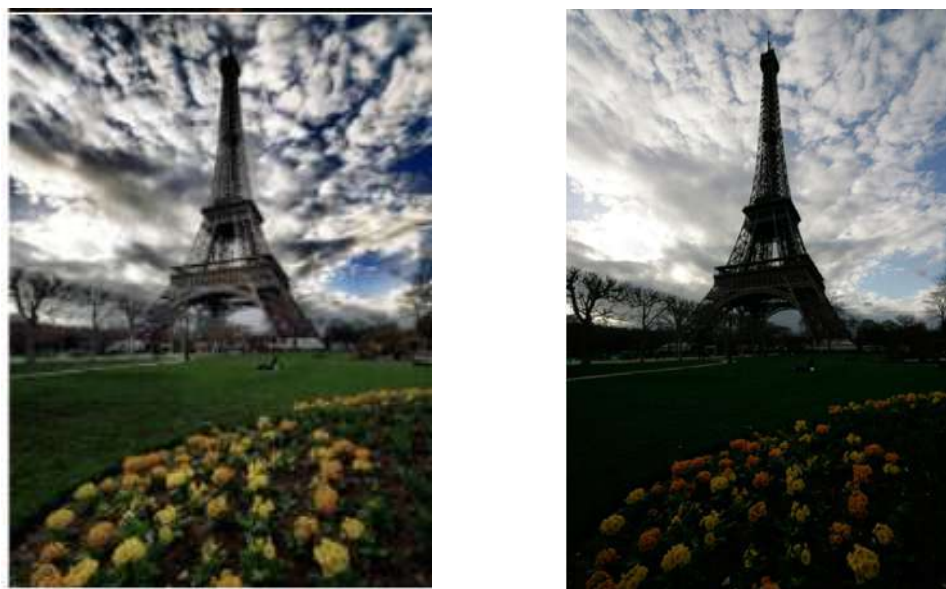


FIGURE 7.7 – Si la régression n'est pas assez robuste, elle peut manquer d'identifier correctement la tour Eiffel en cas de changement de luminosité.

statistique linéaire, malheureusement, les risques d'erreur s'accroissent proportionnellement à mesure que l'on s'éloigne des situations connues. Pour améliorer la robustesse des prédictions, il faut envisager des méthodes non linéaires. Celles-ci peuvent être obtenues en utilisant une fonction de décision non-linéaire.

Dans le perceptron que nous avons envisagé jusqu'à présent, rappelons-le, les réponses sont soit $y = 1$, soit $y = -1$. Ce choix suit une fonction « escalier », aussi appelée fonction de Heaviside, selon que la somme $x_1W_1 + x_2W_2$ soit inférieure ou supérieure à un certain seuil θ . Or il est possible de prendre, pour fonction de décision, une sigmoïde (figure 7.8). Dans ce cas, les valeurs de y vont varier entre -1 et 1 . Il est en outre permis d'utiliser la réponse produite par un premier neurone comme signal incident pour un second neurone. En juxtaposant les neurones les uns les autres, et en utilisant des fonctions de décisions non linéaires, on peut affiner la robustesse des régressions statistiques. On parle alors de réseaux de neurones à couches multiples. Ces réseaux feront l'objet des prochaines sections².

Avant d'entrer dans le vif du sujet, j'aimerais faire remarquer que les

2. Je ne donne ici qu'une caractérisation sommaire des techniques des réseaux de neurones. Le lecteur curieux de connaître plus en détails les approches utilisées avant les années 2000 est invité à consulter le très bon ouvrage vulgarisé d'un de mes professeurs de mathématiques à l'UCLouvain (MICHEL VERLEYSEN, FRANÇOIS BLAYOT, *Les réseaux de neurones artificiels (Que sais-je ?)*, Presses Universitaires de France, 1996). Pour une exposition plus contemporaine, se référer également à STUART RUSSELL, PETER NORVIG, *Artificial Intelligence : A Modern Approach*, Pearson Education.

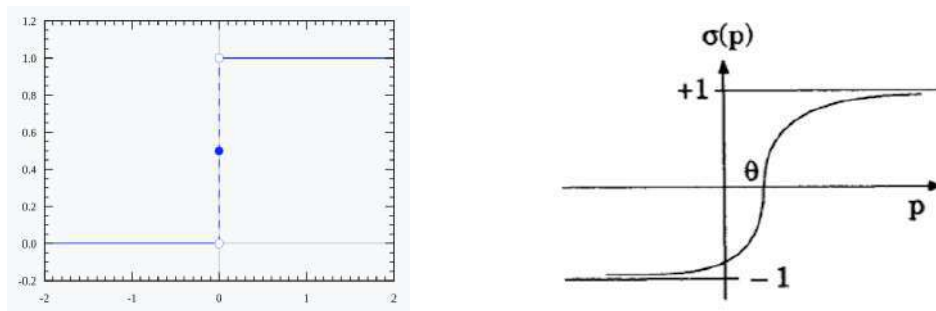


FIGURE 7.8 – La fonction de Heaviside, à gauche, ne peut prendre pour valeur que -1 ou 1 . La fonction sigmoïde au contraire produit des sorties continues entre ces deux points extrêmes.

méthodes les plus élégantes ne sont pas toujours celles qui calculent les paramètres exacts, mais celles qui s'adaptent de façon organique. Les approches computationnalistes, qui assimilent l'esprit humain au cerveau et le cerveau à un ordinateur, en supposant que les décisions neuronales sont le fruit de calculs analytiques similaires à ceux réalisés par un circuit électronique, sont en effet régulièrement battues en brèche par les progrès neuroscientifiques. C'est ce qu'exprime assez éloquemment Francisco Varela dans un entretien avec Herve Kempf³:

La notion d'un ordinateur « neuronal » n'est pas évidente, parce qu'un ordinateur, stricto sensu, c'est un système symbolique. Tout le courant qu'on appelle aujourd'hui le cognitivisme, ou computationnalisme, travaille avec l'hypothèse que le niveau cognitif est autonome. Leur hypothèse est qu'il y a une différence entre les niveaux d'implémentation, entre le soft et le hard. Donc parler de l'ordinateur « neuronal », c'est déjà plonger dans la perplexité la moitié des chercheurs du domaine, pour lesquels ce que fait le cerveau n'est pas directement pertinent pour comprendre la cognition. Si vous parlez à un informaticien, il se moque de savoir sur quel ordinateur il travaille, ce qui l'intéresse, c'est la structure de son programme. De la même façon, ceux qui s'intéressent à la computation en tant que telle ne se préoccupent pas de son support. Ils ne déniaient pas l'intérêt de la biologie, ils disent que ce n'est pas leur domaine. Par ailleurs, il y a énormément de manières de tisser le passage entre l'hypothèse computationnaliste et le cerveau. La question revient à analyser ce qu'est un calcul fondé sur une machine non symbolique, comparable à un cerveau plutôt qu'à un ordinateur.

3. FRANCISCO VARELA, *La Recherche*, No. 308 Avril 1998, p.109-112.

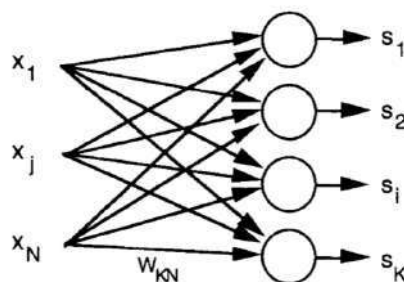


FIGURE 7.9 – Réseau de neurones à une couche.

7.2.2 Réseau à une couche

Nous avons jusqu'à présent envisagé un neurone formel isolé, dont la réponse consiste en une unique décision du type "manger du foin" ou "manger de l'avoine". Or il existe plein d'applications où la réponse à fournir ne peut pas être décrite par une seule variable. L'on peut penser à une intelligence artificielle qui aurait pour vocation de dessiner une image avec 1024×1024 pixels ou à un programme de pilotage aérien automatique prenant en charge l'orientation de l'aileron, du manche, du gouvernail de profondeur. Pour ce genre de situation, les méthodes vues précédemment restent inchangées, mais on introduit, pour chaque nouvelle sortie souhaitée, un nouveau neurone formel. Celui-ci prend en entrée toutes les variables $x_1, x_2, \dots, x_n$ constitutives du problème. Il dispose de même d'un coefficient synaptique spécifique pour chacune de ces entrées.

Les règles d'ajustement des coefficients W peuvent suivre les mêmes approches que celles envisagées précédemment. En présence d'un nombre k de neurones, il suffit d'effectuer le calcul k fois pour les k sorties différentes. Selon l'intérêt de l'application envisagée, des méthodes algébriques ou itératives peuvent être utilisées. Cependant, la puissance de traitement d'une couche de neurones n'est pas différente de celle d'un seul neurone. Il s'agit d'un séparateur linéaire. La seule différence est que la couche neuronale est un ensemble de séparateurs linéaires. Pour résoudre des problèmes qui ne sont pas linéairement séparables, la structure du réseau doit être renforcée, notamment pour permettre la composition des couches.

7.2.3 Réseaux à couches multiples

Il est possible d'ajouter au réseau de neurones plusieurs couches intermédiaires entre les entrées $x_1, x_2, \dots, x_n$ et les sorties $s_1, s_2, \dots, s_k$. Ces couches $C_1, C_2, \dots, C_q$ sont peuplées de neurones qui prennent, en entrée, les sorties de la couche précédente. Chacun de ces neurones produit lui-même une réponse qui servira d'entrée aux neurones de la couche suivante, et ainsi de suite. Notons qu'il n'y a pas d'obligation que ces couches disposent chacune

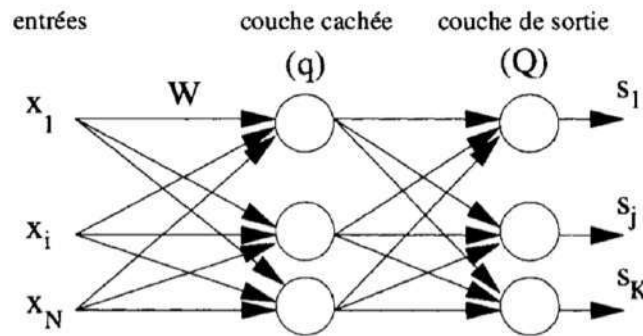


FIGURE 7.10 – Réseau de neurones à couches multiples.

du même nombre de neurones. Il ne faut pas non plus que ces neurones disposent de la même fonction de décision σ . Pour obtenir une généralisation performante du modèle de l'Adaline, il est toutefois judicieux d'associer tous les neurones à une fonction σ non linéaire et différentiable, comme la sigmoïde ou, de façon très similaire, la tangente hyperbolique⁴.

Notons que les neurones des couches intermédiaires ne présentent pas nécessairement le même nombre de neurones que la couche initiale et la couche finale. Le nombre de couches, tout comme la taille de ces couches, est un paramètre libre. Généralement, ceux-ci sont fixés par un programmeur expérimenté, qui détermine, au gré de ses tâtonnements, la meilleure structure. L'on verra plus loin qu'il est également permis de confier cette tâche à un algorithme, qualifié d'algorithme d'auto-apprentissage, ou encore, d'apprentissage non supervisé.

Des coefficients synaptiques sont introduits pour toutes les connexions existantes du réseau. Il y en a donc un associé à chaque arête de la figure ci-contre. Le problème qui se présente consiste à se demander, une fois encore, comment choisir la valeur de ces paramètres? Ce n'est pas une question aisée. Dans les réseaux que nous avons envisagés jusqu'à présent, en effet, l'approche consistait essentiellement à réduire, à l'aide d'une méthode de descente de gradient ou de la méthode des moindres carrés, l'erreur observée entre la sortie qui est produite par le réseau de neurones et la sortie souhaitée. Or, dans le cas d'un réseau de neurones à couches multiples, il semble impossible de dire quelle est l'erreur commise par les neurones des

4. La tangente hyperbolique $\tanh$, définie comme le rapport du cosinus hyperbolique et du sinus hyperbolique, présente une forme très similaire à la fonction sigmoïde. Pour cette raison, elle est souvent utilisée dans les réseaux de neurones. La sigmoïde $\text{sig}(x)$ répond en effet de l'équation $\frac{1}{(1+e^{-x})}$ tandis que $\tanh(x) = \frac{\cosh(x)}{\sinh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^x(1 - e^{-2x})}{e^x(1 + e^{-2x})} = \frac{1 - e^{-2x}}{1 + e^{-2x}} = \frac{2 - (1 + e^{-2x})}{1 + e^{-2x}} = \frac{2}{1 + e^{-2x}} - 1 = 2\text{sig}(2x) - 1$. La tangente hyperbolique est donc deux fois plus escarpée que la fonction sigmoïde. Elle est en outre décalée d'une unité vers le bas.

couches intermédiaires. Les données dont on dispose prescrivent uniquement quelles doivent être, pour un jeu d'entrées $x_1, x_2, \dots, x_n$ donné, les sorties observées à la fin du processus, à hauteur de la dernière couche de neurone. Les opérations qui sont faites dans les couches qui précèdent ne doivent viser qu'à améliorer la sortie finale du réseau.

Pour pallier cette épineuse difficulté, P. Werbos, D. Rumelhart et le Cun ont introduit, et successivement affiné, la méthode connue sous le nom de *rétro-propagation du gradient*^[5]. Cette dernière se fonde sur une approche assez instinctive. Puisque seule la sortie de la couche finale peut être utilisée pour définir une marge d'erreur, il faut associer aux couches intermédiaires des « fonctions objectif » à minimiser. Pour ce faire, supposons, sans perte de généralité et sans souci de notation, que le réseau de neurones comporte q couches de neurones notées $C_1, C_2, \dots, C_q$. On commence par associer à l'avant-dernière couche du réseau (la couche C_{q-1}) une fonction objectif f_{q-1} qui dépend uniquement des poids synaptiques reliant C_{q-1} et C_q . La fonction f_{q-1} doit être telle que, lorsque f_{q-1} est minimum, l'erreur observée sur les sorties de la dernière couche est elle aussi minimum. De cette manière, on a bien un objectif local, pour la couche C_{q-1} , qui coïncide avec l'objectif global concernant les sorties du réseau. Pour doter les autres couches intermédiaires d'un objectif semblable, il suffit alors de répéter la procédure : on associe une fonction f_{q-2} à la couche C_{q-2} , dont le minimum coïncide avec le minimum de f_{q-1} , et ainsi de suite.

L'inconvénient principal de la méthode de rétro-propagation du gradient vient de son temps de calcul très élevé, ce qui induit un apprentissage très long. Pour certaines applications, où le réseau peut être soigneusement optimisé avant son usage, cela n'est pas d'une importance cruciale. Mais s'il est question de modéliser le comportement de la cognition humaine, qui s'adapte en continu aux problèmes rencontrés, on pourra craindre un certain manque de fluidité. Pour contourner cette impasse, il est possible de recourir à d'autres méthodes de minimisation, telles que la méthode du gradient conjugué ou des approches inspirées de Newton-Raphson^[6].

7.3 Deep Learning : Mythe ou réalité ?

Plutôt que de laisser entre les mains d'un programmeur expérimenté le choix des paramètres intrinsèques du réseau de neurones, tels que le nombre de couches, le nombre de neurones par couche, le type de fonction à minimiser ou la méthode de minimisation utilisée, ce choix peut être confié à un algorithme. Cette voie n'exempte pas le recours au jugement humain en

5. D. RUMELHART, D. MCCLELLAND, *Parallel Distributed Processing. Exploration in the microstructure of cognition, vol I*. Cambridge, MIT Press.

6. Ces méthodes d'optimisation consistent *grosso modo* à calculer la pente de la fonction et à chercher où elle s'annule. Elles étaient assez populaires au début de l'intelligence artificielle, mais ne sont plus beaucoup utilisées.

amont du processus, mais elle donne à la machine une plus grande autonomie. Cette dernière est même susceptible d'évoluer itérativement vers une forme non planifiée par le concepteur : son réseau peut s'accroître ou se restreindre de façon organique, ses poids synaptiques peuvent être ajustés dès l'injection de nouvelles données, et sa structure, pour être bref, est capable d'évoluer comme un simulacre d'être vivant.

Un des premiers réseaux de neurones évolutifs envisagés dans la littérature est celui de Reilly et Cooper^[7], baptisé le Restricted Coulomb Energy (RCE). Contrairement aux algorithmes envisagés jusqu'à présent, qui visent à séparer linéairement ou non linéairement des catégories de données, le RCE a pour rôle de déterminer des régions d'influence. Ce que j'entends par là, c'est que l'algorithme ne vise pas à discriminer toutes les situations dans lesquelles la réponse optimale est 1 (manger du foin) ou -1 (manger de l'avoine), mais à dessiner une sorte de zone de confiance autour des cas où la réponse est connue.

Pour donner une illustration géométrique, on pourrait dire que la tâche usuelle d'un réseau de neurones consiste à découper des zones, ou catégories, au sein de l'espace des données possibles. de données a autant d'axes que le problème ne comporte d'entrées $x_1, x_2, \dots$. Dans le cas de l'âne de Buridan, par exemple, il s'agit d'un espace à deux dimensions avec en abscisse le degré d'appétit et en ordonnée le degré d'enrhumement, mais dans des cas plus complexes, l'espace peut comporter des centaines de dimensions. En calculant les poids $W_1, W_2, \dots$, le réseau de neurones détermine des frontières au sein de cet espace. Dans le cas d'un plan à deux dimensions, il s'agit de droites, mais dans le cas plus général, on préférera parler d'hyperplans. Si elles sont bien construites, ces frontières doivent être telles que, lorsque l'on se trouve d'un côté, la sortie est $+1$; tandis que de l'autre côté de la frontière, la sortie est -1 . Si le réseau admet plus que $+1$ et -1 comme réponse, il peut y avoir des frontières plus complexes et plus nombreuses. Il n'y a cependant aucune région de l'espace qui ne soit pas associée à une réponse de la part du réseau de neurones.

Or, dans le cas de l'algorithme RCE, on ne cherche pas à découper tout l'espace des données possibles en catégories distinctes ; mais on part uniquement des points connus (les données d'apprentissage) et on dessine, autour de chacun d'eux, des zones dans lesquelles il est probable que la réponse correcte soit la même que celle associée au point de départ. Si l'âne de Buridan, pour reprendre notre exemple canonique, mangeait de l'avoine lorsque $x_1 = 0.8$ et $x_2 = 0.2$, l'algorithme peut décider que, dans un petit rayon autour du point $(0.8, 0.2)$, l'âne doit manger de l'avoine. Cette approche est dite constructive, car les régions d'influence ne sont pas obtenues en découpant le plan initial avec des hyperplans, mais en construisant des zones de

7. D. REILLY, L. COOPER, C. ELBAUM, *A neural model for category learning*, Biological Cybernetics, 45, 35-41, 1982.

plus en plus grandes autour des points connus.

D'un point de vue procédural, l'algorithme commence par dessiner un rayon arbitraire (fixé par les paramètres de RCE) autour des données qu'il détient. Si deux zones entrent en collision, il y a deux possibilités. Soit ces deux zones sont associées à la même réponse ($s = 1$ = "manger du foin"), soit elles sont associées à des réponses distinctes ($s = 1$ = "manger du foin" et $s = -1$ = "manger de l'avoine"). Dans le premier cas de figure, l'algorithme reçoit une confirmation de son classement et il ne change pas. Dans le second cas de figure, il va diminuer les rayons d'influence autour des points envisagés de façon à ce qu'ils n'entrent plus en collision. L'algorithme procède ainsi itérativement, jusqu'à ce qu'il n'y ait plus aucune collision entre les zones associées à des réponses distinctes.

On comprendra que ce type d'algorithme ne dira rien sur les points trop éloignés des données de départ. Dans le pire scénario, il est même possible que le nombre de régions d'influence soit égal au nombre de données d'apprentissage. Cela se produit lorsque le rayon de départ est trop petit, ou lorsque les catégories sont si imbriquées les unes dans les autres, et de façon si complexe, que la méthode RCE ne parvient pas à généraliser ses conclusions au-delà d'une portée très restreinte. Dans les deux cas de figure, il en résulte une incapacité, pour le réseau de neurones résultant, de catégoriser des entrées qui sont un tant soit peu éloignées des données d'apprentissage. Pour le dire en d'autres termes, le réseau est incapable de s'adapter aux situations nouvelles et performera de façon très peu satisfaisante.

Pour éviter cet effet, que l'on appelle le sur-échantillonnage, il peut être judicieux d'augmenter ou de diminuer stratégiquement le nombre de couches et le nombre de neurones par couche. En augmentant le nombre de couches intermédiaires, et la taille de ces couches, on s'autorise à construire des frontières non linéaires entre les catégories, ce qui est particulièrement pratique en présence de catégories qui sont fort imbriquées les unes dans les autres. Mais si la précision est trop fine, les frontières construites n'embrasseront pas les frontières réelles du problème.

J'aimerais donner à ce problème une analogie inspirée de l'analyse numérique. Lorsque l'on dispose de points dans l'espace et que l'on cherche à les relier au moyen d'un polynôme du type $a_0 + a_1x + a_2x^2 + \dots$ (où $a_0, a_1, \dots \in \mathbb{R}$), le choix se pose de savoir jusqu'à quel degré polynomial on souhaite monter ? Va-t-on se contenter d'une droite ($a_0 + a_1x$) ? D'une parabole ($a_0 + a_2x^2$) ? D'une courbe du troisième degré ($a_0 + a_2x^2 + a_3x^3$) ? On pourrait croire, naïvement, qu'au plus grand est le degré du polynôme, au plus grande est la précision, mais il n'en va pas toujours ainsi. Le graphique ci-dessous illustre ce qui se passe lorsque l'on prend un degré polynomial trop petit ou trop grand pour relier les points les uns aux autres. On appelle phénomène de Runge les oscillations extrêmes qui apparaissent en augmentant excessivement le degré du polynôme.

Le nombre de coefficients synaptiques d'un réseau de neurones est équi-

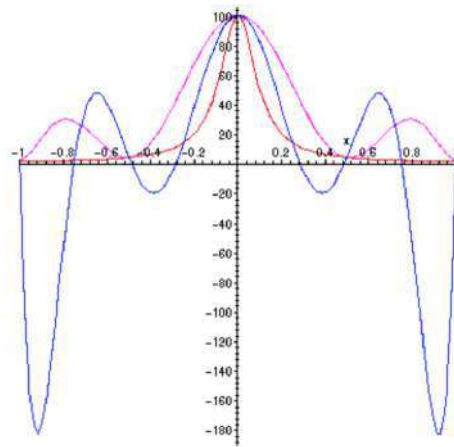


FIGURE 7.11 – Illustration du phénomène de Runge. La ligne en rouge est le graphique de la fonction $f(x) = \frac{1}{x^2 + 0.01}$. Les lignes bleues sont les interpolations de quelques points choisis de la fonction f . On observe que les polynômes de plus hauts degrés présentent une déviation plus importante par rapport à la courbe rouge (Bibm@th.net).

valent, dans notre analogie, au degré du polynôme d'interpolation. Si le réseau comporte trop de connexions synaptiques, les frontières dessinées par un réseau de neurones peuvent ne pas épouser correctement le comportement attendu. Dans le cas précis du RCE, il en résultera des zones d'influence beaucoup trop petites, incapables de généraliser les réponses connues à des cas nouveaux. Dans le pire scénario, il peut arriver que les performances en soient si fortement affectées que le réseau sera incapable de dire quoi que ce soit de pertinent en dehors des données d'apprentissage. Il est toutefois important de remarquer que, pour un nombre de synapses identiques, il vaut mieux avoir un réseau à plusieurs couches qu'à une seule couche.

Pour cette raison, l'algorithme RCE va être doté d'un procédé consistant à augmenter progressivement le nombre de couches, puis à expurger le réseau résultant des neurones inutiles, et ce, itérativement, dans l'espoir d'optimiser les rayons des zones d'influence. La suppression des neurones excédentaires peut se faire pendant l'apprentissage, mais elle peut aussi avoir lieu après coup. Dans les deux cas de figure, l'élagage demande de déterminer quels sont les neurones les plus utiles au réseau et quels sont ceux qui peuvent être retirés sans trop de peine. Pour ce faire, on peut par exemple enlever les neurones un à un et comparer quel effet cela produit sur les sorties du réseau ? L'erreur moyenne augmente-t-elle d'une marge significative ? Cette comparaison peut être faite en retirant un seul neurone à la fois ou en retirant plusieurs d'entre eux. On essaie de garder le sous-réseau (c'est-à-dire une partition du réseau initial) dont la fonction d'erreur reste au plus proche

du réseau initial (et sans dépasser certaines bornes concernant la baisse de performance). Il existe des méthodes plus développées, telles que l'OBD (Optimal Brain Damage^[8]), l'OBS (Optimal Brain Surgeon^[9]) ou la SSM (Statistical Stepwise Method^[10]).

7.3.1 Programmation génétique

Nous avons décrit une manière de faire évoluer le code au fur et à mesure de son apprentissage. Il existe cependant d'autres approches dignes d'être mentionnées. Une méthode célèbre, appelée la programmation génétique, peut être utilisée aussi bien pour faire évoluer un réseau de neurones qu'un algorithme de type top-down. Cette méthode est introduite en 1958 par Friedberg^[11], avant de recevoir des raffinements de la part de Cramer, ou de Koza. Elle est aujourd'hui envisagée, notamment, pour la mise au point d'agents moraux artificiels^[12].

Comme son nom l'indique, la programmation génétique emprunte à la théorie de Darwin ses lettres de noblesse. Elle consiste à générer une population d'algorithmes aléatoires et à ne garder que ceux qui sont les plus performants face au problème que l'on cherche à résoudre. À l'issue de cette mise en compétition, appelée un tournoi, on retient les algorithmes ayant affiché les meilleures performances. On va alors brasser leurs paramètres de façon aléatoire pour obtenir une nouvelle population de concurrents. En répétant cette méthode de façon itérative, on souhaite converger vers un code de plus en plus robuste. Dans le langage biologique, on dira qu'on espère augmenter l'adaptation du code à chaque nouvelle génération.

Supposons que nous souhaitions créer une intelligence artificielle pour le jeu d'échecs. Il faut préalablement se munir d'un code de départ. Il peut s'agir d'instructions de type top-down, d'un réseau de neurones ou d'une approche par force brute. L'essentiel est que certains paramètres de ce code puissent faire l'objet d'une évolution. Ces paramètres peuvent être dédiés à l'évaluation des positions, comme la valeur associée à chaque type de pièce, l'importance donnée à la sécurité du roi, au développement des pièces, au contrôle de l'échiquier ou à la structure de pions ; mais il peut aussi s'agir de paramètres relevant du code proprement dit, comme le nombre et la taille des couches du réseau de neurones (s'il s'agit d'un réseau de neurones) ou

8. YANN LE CUN, JOHN S. DENKER et SARA A. SOLLA, *Optimal Brain Damage*, AT&T Bell Laboratories, Holmdel, N. J. 07733 .

9. BABAK HASSIBI, DAVID G. STORK, *Second order derivatives for network pruning : Optimal Brain Surgeon*, Ricoh California Research Center, Stanford, CA 94305.

10. N. FNAIECH, F. FNAIECH, B. W. JERVIS, M. CHERIET, *The Combined Statistical Stepwise and Iterative Neural Network Pruning Algorithm*, November 2009, Intelligent Automaton & Soft Computing 15(4) : 573-589).

11. Friedberg, R.M. (1958) A Learning Machine : Part I. IBM Journal of Research and Development, 2, 2-13.

12. Muntean & Howard, 2014

la profondeur de l'analyse en nombre de tours.

À chacun de ces paramètres doit être associée une fonction de probabilité raisonnable, qui assigne au paramètre une valeur aléatoire dans l'intervalle $[a, b]$. Cette fonction doit être assez générale pour ne pas exclure de possibilités intéressantes, mais il faut la contraindre quelque peu pour accélérer le processus de convergence. L'importance donnée à la valeur de chaque pièce, par exemple, pourra être prise dans $[0, 20]$, avec une valeur fixée infinie pour le roi¹³. Il est important de permettre des valeurs nulles, car il se peut que l'algorithme, au fil des itérations, converge vers une solution où certains paramètres s'avèrent inutiles. Cela se produit, par exemple, lorsque l'on observe que la valeur associée à chaque pièce du jeu tend vers une même valeur, ou que les valeurs des pièces ne comptent presque pour rien dans l'évaluation finale d'une position. Toutes les fois que cela se produit, on peut retirer sans crainte le paramètre.

Lorsque les paramètres sont fixés, on va organiser un grand nombre de parties afin de tester les algorithmes les uns contre les autres. Un système de points doit être mis au point afin de déterminer les algorithmes les plus performants. Pour le jeu d'échecs, on attribue généralement un point en cas de victoire et un demi-point en cas d'égalité. Il existe également des systèmes permettant de départager les égalités (Buchholz, Buchholz tronqué et la somme cumulée¹⁴). Cela permet donc de dresser la liste des algorithmes ayant eu les meilleures performances. On conserve les meilleurs d'entre eux, par exemple 10% des meilleurs, ou 5%. Ce pourcentage doit être assez élevé pour autoriser une certaine diversité dans le profil des vainqueurs. Il se peut en effet qu'il y ait plus d'une seule stratégie gagnante : certains jeux de paramètres, entre les vainqueurs, seront sensiblement différents.

Il convient pour terminer de mélanger aléatoirement les paramètres des joueurs vainqueurs afin de générer une nouvelle génération de compétiteurs. L'approche la plus naïve consiste à appairer aléatoirement les programmes entre eux et à retenir, pour chaque appariement, la moyenne de leurs paramètres. Si l'algorithme *A* possède la valeur 0,736 sous le paramètre « sécurité du roi » et que l'algorithme *B* possède la valeur 0,112, leur « des-

13. Dans la plupart des langages de programmation, il est possible d'attribuer aux variables numériques une valeur dite "infinie". En pratique, il s'agit de la plus haute valeur qui peut être prise par un entier. Dans le langage Java, cette valeur vaut 2 147 483 647. Si l'on ajoute le nombre un à cette valeur, on retourne à zéro. Il existe une convention permettant de distinguer lorsque l'entier représente la plus grande valeur ou lorsqu'il représente strictement l'infini.

14. Le score de Buchholz est la somme des points marqués par les adversaires rencontrés. Buchholz encourage donc le joueur ayant joué contre des adversaires plus performants. Buchholz tronqué suit le même principe, mais en retirant de cette somme le plus faible des adversaires. La somme cumulée, pour finir, est la somme des points obtenus à l'issue de chaque ronde. Selon ce principe, les points marqués en début de tournoi ont plus d'importance que ceux marqués par la suite. C'est un indicateur qui vise à prévenir une stratégie pernicieuse : celle qui consiste à perdre volontairement au début du tournoi pour être apparié à des adversaires plus faibles.

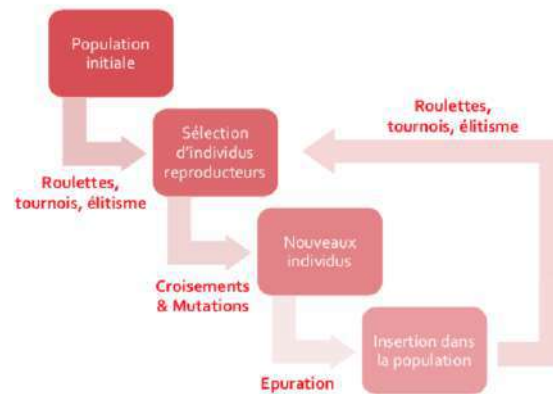


FIGURE 7.12 – La programmation génétique est un processus itératif consistant à optimiser, par sélection-élimination, une population de paramètres aléatoires.

ependant » sera un algorithme ayant pour valeur 0,424. Il peut toutefois être insatisfaisant de prendre la moyenne, car cela peut s'apparenter à un nivellement vers le bas. Des jeux de paramètres correspondants à des stratégies efficaces, quoique peu représentatives au sein des gagnants, risquent de disparaître pour la simple et bonne idée que leurs paramètres sont moyennés avec les paramètres d'une stratégie plus dominante. C'est pourquoi il peut être préférable de sélectionner aléatoirement une des deux valeurs, plutôt que de prendre leur moyenne. Une troisième voie, plus subtile encore, consiste à les moyennner au travers d'une fonction sigmoïde. Chacun des deux paramètres compte alors comme l'entrée d'un neurone formel. La sortie de ce neurone est le nouveau paramètre à utiliser. Il existe, comme l'on s'en doute, de nombreuses façons de paramétrer ce neurone formel.

J'espère avoir montré, au moyen de ce bref discours sur la programmation génétique, qu'il existe des méthodes de programmation bottom-up qui ne reposent pas sur des réseaux de neurones. Par ailleurs, je m'estime satisfait si le lecteur en ressort avec une idée, même approximative, de la façon dont les réseaux de neurones peuvent être supportés par la sélection génétique, ou comment les réseaux de neurones, en revanche, peuvent venir renforcer cette dernière. La plupart des codes performants qui se rencontrent dans la littérature mélangent astucieusement les différentes recettes. Ce sont des hybridations dignes de l'authentique chimère de la mythologie, mais dont leur efficacité, elle, est loin d'être chimérique.

7.3.2 Apprentissage non supervisé

Les méthodes que nous avons vues jusqu'à présent consistent à remettre, entre les mains de la machine, le gouvernail du navire de la régression statis-

tique, mais tout en continuant à superviser sa conduite. L'être humain fixe le cap à suivre. C'est lui qui choisit les données d'apprentissage et la façon de les interpréter, en prescrivant quelles sont conduites à imiter et celles à éviter.

En termes informatiques, on dit que le programmeur « étiquette » les données. Certaines entrées sont considérées « bonnes » et d'autres « mauvaises ». Dans le cas de l'âne de Buridan, par exemple, c'est l'être humain qui prescrit quels ont été les bons choix par le passé, de façon à inciter le neurone formel à reproduire les décisions avantageuses et à s'écarter des autres. Dans un logiciel d'échecs, pareillement, la base de données est constituée de situations de jeu (des configurations de pièces sur l'échiquier), dont certaines sont jugées favorables et d'autres défavorables. Dans un programme visant à reconnaître des formes géométriques, pour finir sur un exemple quelque peu différent, on prescrit d'avance les formes qui doivent être identifiées : lettres, carrés, triangles, etc.

Le jugement permettant d'étiqueter les données repose sur des heuristiques, ou principes d'évaluation intuitifs, conçus par l'être humain. La plus connue de ces heuristiques, aux échecs, est la fameuse règle du pouce, qui dit qu'une dame vaut dix points, une tour cinq points, un fou trois, un cavalier deux et demi, un pion un point¹⁵. Ces heuristiques présentent ceci de commode qu'elles permettent de quantifier *a priori* la valeur des situations de jeu rencontrées. Il n'est toutefois pas aisé de s'entendre sur ces heuristiques ; et elles ne sont pas valides dans tous les contextes. Les prouesses échiquéennes regorgent d'exemples où un grand maître sacrifie du matériel pour obtenir un avantage positionnel substantiel.

C'est pourquoi une nouvelle génération de réseaux de neurones, dits non supervisés, partent du principe que les catégories ne doivent pas être déterminées préalablement par le concepteur humain. Qu'est-ce que j'entends par là exactement ? Supposons que l'on souhaite créer un algorithme capable de reconnaître des chiffres écrits de façon manuscrite. Dans l'approche standard, on doit donner à la machine un exemple prototypique de chaque chiffre existant : un exemple de 0, de 1, de 2, etc. La machine s'appuie sur ses exemples pour classer tous les nouveaux caractères rencontrés. Elle est très efficace pour identifier des chiffres connus, mais ne sera pas capable de créer ses propres catégories.

Dans une approche non supervisée, on ne donne aucun prototype de départ à l'algorithme ; on ne lui indique même pas combien de chiffres différents comprend l'alphabet. On la laisse comparer statistiquement les lettres manuscrites les unes aux autres pour identifier des similarités par elle-même.

15. On en trouve une formulation chez Alan Turing, qui donnait neuf points pour la dame. À cette époque, Turing devait calculer lui-même les réponses de l'algorithme sur une feuille de papier et cela lui prenait trente minutes par mouvement. Voir notamment A. M. TURING, *Intelligent Machinery*, Collected Works of A.M. Turing, Mechanical Intelligence, p.127.

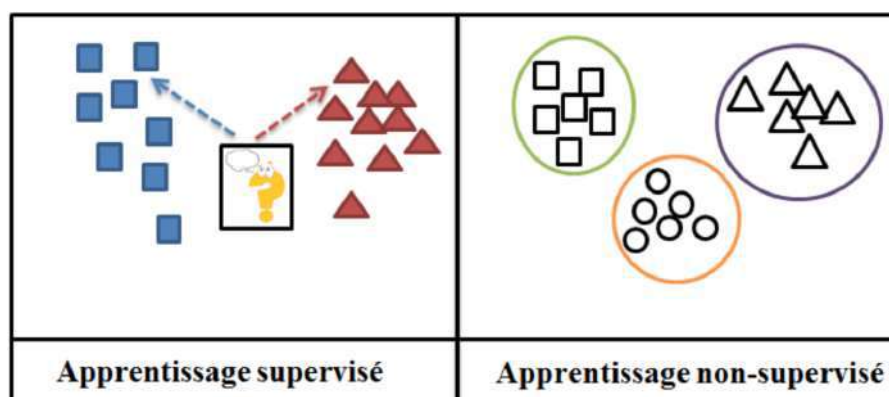


FIGURE 7.13 – Dans l'apprentissage non supervisé, la machine trie elle-même les données en les rangeant dans différentes classes, de sorte à maximiser l'indépendance statistique des groupes obtenus.

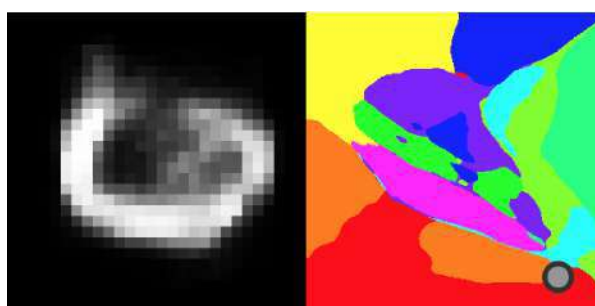


FIGURE 7.14 – Cet algorithme non supervisé a pour tâche de reconnaître des chiffres manuscrits sans qu'on lui ait donné un prototype de chaque chiffre existant. On observe de bonnes performances, même si certaines lettres proches du 0 sont classées comme des 6 (source : <https://n8python.github.io/mnistLatentSpace/>).

La catégorisation émerge ainsi en comparant, deux à deux, les input disponibles. Cette différence de méthode est analogue à celle qui existe entre un enfant qui apprend à lire sur les bancs de l'école, sous les sages conseils du professeur, et l'enfant qui apprend à lire par lui-même.

L'avantage de l'apprentissage non supervisé, c'est qu'il permet à la machine de créer des classes de données qui ne seraient pas venues à l'idée de l'être humain. Face à un alphabet inconnu, par exemple, la machine peut d'elle-même identifier de nouvelles lettres. Le désavantage, en revanche, c'est qu'il est possible que des variations de la même lettre soient cataloguées dans des classes différentes. Dans le pire scénario, on peut imaginer que la machine distingue tout autant de lettres qu'il n'y a de styles d'écriture : le A en écriture cursive sera distingué du A de l'écriture scripte, etc.

La méthode de catégorisation la plus souvent utilisée est le clustering, qui consiste à trier les données disponibles dans des classes de similarité. Cela peut être réalisé en choisissant la partition qui maximise l'indépendance statistique des classes les unes vis-à-vis des autres. Cette indépendance statistique peut être mesurée à l'aide de plusieurs indicateurs : la divergence de Kulbak-Leibler, la PCA, l'IPC, etc. Je ne m'étendrai pas sur ces différentes méthodes, mais j'aimerais que le lecteur comprenne l'autonomie qui échoit à la machine dans l'apprentissage non supervisé.

Il peut arriver que les classes obtenues reposent sur des différences statistiques qui n'ont pas de signification dans le langage humain. Une machine visant à la reconnaissance des caractères pourrait tenir compte, par exemple, de la couleur des caractères, de la variation d'orientation ou de la luminosité de la page, notions dont l'esprit humain fait abstraction lors de la lecture. Or comme le programmeur ne sait pas d'avance comment l'algorithme va se comporter, il peut en résulter un problème de traçabilité : il est de moins en moins aisé de dire pourquoi et comment le réseau de neurones est arrivé à telle ou telle conclusion. Cette opacité intrinsèque est probablement l'inconvénient le plus significatif des méthodes bottom-up.

Notons pour terminer qu'il existe des stratégies d'apprentissage semi-supervisées où certaines données seulement sont étiquetées. Cela permet de s'assurer que les classes voulues apparaissent à la fin du processus, tout en permettant à la machine de construire ses propres schémas.

7.4 Critique épistémologique : Entre instinct et raisonnement

Il est temps de synthétiser les résultats obtenus par le paradigme de programmation bottom-up et de mettre en relief certaines faiblesses épistémologiques. Notre petite trajectoire au sein des réseaux de neurones nous a permis d'identifier deux stratégies générales de Deep Learning : l'apprentissage supervisé et l'apprentissage non supervisé. Dans le premier cas de figure, les catégories dans lesquelles la machine doit classer les données sont définies *a priori* par le programmeur. Ce dernier stipule, par exemple, que les images à analyser représentent des formes géométriques de types « cercle », « triangle » ou « carré ». Il donne en outre des exemples typiques de cercles, de triangles et de carrés sur lesquels l'algorithme peut s'appuyer pour rendre sa décision.

Dans l'apprentissage non supervisé, au contraire, le réseau de neurones est confronté à des données brutes et il doit, par lui-même, créer des concepts. Ces derniers émergent en regroupant les données de l'échantillon dans des classes statistiquement indépendantes les unes des autres. Face à des figures géométriques, la machine pourrait par exemple regrouper tous les triangles entre eux, mais pour le même prix, elle pourrait également trier les formes d'après leur couleur, en mettant les rouges d'un côté et les verts de l'autre. Il n'y a aucune garantie que le classement résultant de cette approche épouse

les concepts envisagés par le programmeur. Il se peut que les triangles de grande taille, par exemple, soient associés à une autre classe que les triangles de petite taille. En procédant de la sorte, l'algorithme ne produit pas de critère rationnel permettant d'identifier un triangle ou un carré par leurs définitions. Il ne fait rien que réunir les objets qui présentent entre eux certaines similarités, conformément à l'adage d'après lequel « ce qui se ressemble s'assemble. » Il arrive même que la classification se fasse d'après des similarités qui ne sont pas apparentes pour un être humain.

L'un des problèmes fondamentaux des réseaux de neurones consiste donc à créer une liaison entre les concepts statistiques générés sur base du Deep Learning et les concepts rationnels définis à l'aide de critères logiques préalables, comme le fait de « posséder quatre angles droits » ou « d'avoir trois angles dont la somme fait 180 degrés ». Ce problème que j'évoque demande de fournir ce que Platon appelait un « jugement muni de justification ». Il ne s'agit pas uniquement de classer les données en mettant dans une même case les objets qui se ressemblent les uns les autres ; il faut encore exprimer, sous la forme d'une règle vérifiable et opérationnelle, les points communs que partagent les éléments d'une même classe. Cette règle doit être suffisamment générale pour que tous les représentants de cette classe la satisfassent, mais elle doit être assez exclusive pour que les membres des autres classes ne la satisfassent pas.

Or en l'état, la programmation bottom-up présente un manque de traçabilité. Les réponses qu'elle apporte au problème de classification ne sont pas faciles à expliquer. Le programmeur lui-même ne comprend pas toujours nécessairement la raison pour laquelle l'algorithme dit d'une personne qu'il voit sur une photo qu'il s'agit d'un être humain ou d'un gorille.

7.4.1 Le Talon d'Achille des réseaux de neurones

Pour illustrer cette difficulté que j'évoque à l'aide d'un cas d'école, supposons qu'un réseau de neurones ait pour tâche d'automatiser l'ouverture des portes d'une station de métro, qui était jusque-là contrôlée par un agent humain¹⁶. On voudrait que la machine soit capable d'identifier le nombre de personnes présentes dans la station : en situation normale, elle ne doit ouvrir qu'une seule porte de sortie, mais en cas de forte affluence, un second sas doit être autorisé. Pour entraîner l'algorithme, on lui fournit toutes les vidéos prises par les caméras de surveillance de la station de métro aux cours des dix dernières années. On suppose qu'avec une telle masse de données, il sera en mesure de déterminer par lui-même quelles sont les situations où le sas doit être ouvert. Après cet entraînement, on teste le réseau de neurones en situation réelle et on observe des résultats concluants. On licencie

16. J'ai plusieurs fois lu ou entendu cet exemple dans des expositions pédagogiques, avec quelques variantes. Il s'agit probablement d'un cas fictif.

alors l'agent humain qui s'était chargé jusque là de cette mission certes laborieuse et peu gratifiante — ce qui permet aussi, admettons-le, de réaliser des économies pour le Trésor public.

Or voilà que, un mois plus tard, commencent les Jeux Olympiques. La station de métro est envahie par la foule au beau milieu de l'après-midi, mais l'algorithme omet d'ouvrir le sas supplémentaire. Les gens se bousculent, ils s'écrasent les uns les autres, des accidents sont déclarés. Une enquête est réalisée, mais ni le chef de service, ni les programmeurs de l'algorithme, ni même l'agent humain qui était en fonction jusque là pour ouvrir les portes, ne peuvent dire quelle est la raison de ce comportement inattendu de la part de la machine. C'est alors qu'on découvre que l'algorithme, plutôt que de prendre en compte le nombre de personnes dans la station, se contente d'ouvrir le sas pendant les heures de pointe, entre 7h et 9h et entre 16h et 18h. Il a en effet observé une forte corrélation statistique entre ces heures de la journée et l'ouverture du sas additionnel au cours des dix dernières années. Les catégories statistiques que l'algorithme a forgées pour résoudre le problème n'utilisent pas les critères voulus par ses programmeurs ! La machine s'est appuyée d'une variable (l'heure) qui ne devait pas figurer dans la résolution du problème, mais que les données d'entraînement comportaient néanmoins (les vidéos des caméras de surveillance présentant cette information).

Ce que cet exemple cherche à illustrer, c'est que les raisonnements des réseaux de neurones ne sont pas tout à fait traçables. On ne peut pas expliquer précisément quelles variables sont utilisées par le programme, quelles corrélations sont effectuées ou pourquoi une certaine décision est privilégiée par rapport à une autre. C'est d'ailleurs ce manque de traçabilité qui est à l'origine d'un grand nombre de problèmes éthiques soulevés par l'usage de la programmation bottom-up. J'ai donné quelques perspectives à ce sujet dans le chapitre , au moment de parler des biais algorithmiques. On se souviendra notamment du programme chargé de déterminer si un individu en procédure pénale était, oui ou non, susceptible de récidiver. Une femme noire pouvait recevoir des notes négatives pour des délits pourtant mineures, car des associations statistiques indésirées sont tissées entre l'ethnicité et le casier judiciaire (même lorsque l'ethnicité n'est pas explicitement fournie à l'algorithme).

Cette opacité dont je parle fait qu'on ne peut pas voir les réseaux de neurones comme une horloge mécanique, dont les rouages seraient visibles à l'œil nu, et où rien ne se produirait en l'absence d'une cause identifiable. Il faut bien plutôt y voir une boîte noire, qui prend des inputs et produit des outputs d'après des lois inconnues (ou partiellement connues). Comme l'on est incapable de tester les performances de cette boîte dans toutes les situations possibles, on se contente de suppositions plus ou moins raisonnables. Lorsque les outputs observés lors des tests officiels de l'algorithme sont satisfaisants, on estime que l'algorithme exécute la fonction voulue, ou qu'il est suffisamment proche de cette fonction. Si, par exemple, notre boîte

noire produit toujours la division euclidienne a/b des nombres a et b qu'on lui soumet, on affirmera qu'elle exécute la fonction de division euclidienne. Il n'y a toutefois aucune certitude que le comportement se reproduira avec des paires de nombres a et n différentes de celles utilisées pour tester l'algorithme.

Il est important de remarquer que ce problème que je pointe du doigt n'est pas le propre de la programmation bottom-up, mais qu'il est également présent, quoique de façon moins dramatique, dans la programmation top-down. Dans cette dernière, en effet, on donne les instructions précises et contraignantes au programme. On lui dit par exemple : « ouvre la porte lorsque tel ou tel critère est observé ». Il y a donc une traçabilité quant à ses règles ; l'inspection des lignes de code permet généralement de prédire le comportement de la machine. En dépit de cette traçabilité, on ne peut pas être certain que le programme soit écrit correctement. On a vu dans le chapitre 5.3 que le problème consistant à vérifier la sémantique d'un logiciel était en effet indécidable (par le théorème de Rice) ; ou autrement dit, qu'il n'existait pas de procédure automatique permettant de déterminer en un temps raisonnable à tous les coups si un programme fait bien ce que l'on souhaite qu'il fasse¹⁷.

La différence essentielle entre l'opacité des réseaux de neurones et l'indécidabilité concernant les propriétés sémantiques des algorithmes top-down n'est qu'une différence de degré : si le programme top-down est vérifiable par un humain en utilisant les heuristiques appropriées, le programme bottom-up résiste plus farouchement à la vérification. Tout se passe comme si les réseaux de neurones avaient pris le parti d'embrasser pleinement l'idée que « les algorithmes, c'est de la magie noire ». On les fait tourner et on voit ce qu'il se passe. Il en découle des résultats mitigés. D'une part, on obtient des machines plus flexibles, plus adaptatives, plus performantes face à de nombreuses nouvelles tâches. D'autre part, on le paie au prix d'une incapacité fondamentale de comprendre le mécanisme interne de l'algorithme et d'en prédire les réponses avec un haut degré de certitude. Tout se passe, en grossissant le trait, comme si l'on renonçait à faire usage du jugement humain en amont du processus — à la phase initiale de *design*. Ce jugement ne peut plus intervenir qu'en aval, au moment d'évaluer les résultats, de les interpréter, de les contextualiser.

Certains chercheurs en intelligence artificielle ont très tôt affirmé qu'il était nécessaire de combiner efficacement l'approche bottom-up et l'approche

17. Je dis, en un temps raisonnable, car le théorème de Rice n'est valide qu'en présence d'algorithmes présentant un espace de stockage infini. Dans la pratique, ce n'est pas le cas, si bien que le théorème peut être contourné. Le temps nécessaire pour corriger la sémantique d'un programme ayant une base de stockage n donnée augmente toutefois de façon exponentielle avec n . Cela veut dire que pour des programmes un peu sophistiqués, il devient très vite difficile d'utiliser une procédure de vérification automatique qui fonctionne dans tous les cas de figure.

top-down. Leslie Valiant, par exemple, s'intéresse dès 2008 à des architectures hybrides^[18]. Des stratégies analogues sont entreprises par les tenants de la programmation neurosymbolique, qui allient les techniques des réseaux de neurones à des stratégies typiquement inspirées des systèmes experts. La raison de cet engouement vient notamment de la théorie de Daniel Kahneman, prix Nobel d'économie de 2002, qui insiste sur l'existence, dans le chef de l'esprit humain, de deux systèmes de traitement de l'information : (a) un système automatique, rapide, statistique, qui ressemble au fonctionnement des réseaux de neurones et (b) un système conscient, lent, rationnel, qui s'apparente davantage à la programmation top-down. Ces deux systèmes de Kahneman ne mobilisent pas les mêmes régions neuronales et ne répondent pas des mêmes principes, mais ils sont tous deux importants pour la prise de décision humaine. Dans son livre intitulé *Système 1 et Système 2 : les deux vitesses de la pensée*, Kahneman donne la description suivante^[19] :

Le Système 1 opère de manière automatique et rapide, avec peu d'effort et sans s'accompagner du sentiment d'exercer un contrôle volontaire. Le Système 2 alloue les ressources attentionnelles aux activités mentales exigeantes, y compris les calculs complexes. Les opérations du Système 2 sont souvent associées à l'expérience subjective d'exercer son libre arbitre et de faire un effort de concentration.

Tout au long de cette thèse, j'ai tenté de montrer qu'il existait peut-être un troisième système, situé au milieu des deux premiers. Ce Système 3, que la philosophie appelle traditionnellement du nom de *l'Entendement*, est une opération certes consciente, mais pas totalement justifiable. Elle consiste à penser les concepts d'une façon qui ne soit ni statistique ni rationnelle, mais intuitive. Cette intuition se fait via l'élaboration d'une image mentale, un schème. Cette image est certes inspirée par des représentations concrètes, mais elle désigne quelque chose qui dépasse toutes les représentations possibles. Qu'est-ce que je veux dire par là ? Lorsque vous vous représentez l'idée d'un triangle ou d'un arbre, vous voyez sans doute un triangle isocèle ou un chêne, mais cette image mentale que vous produisez dans votre esprit ne désigne pas un triangle isocèle ou un chêne en particulier ; elle désigne l'idée abstraite du triangle ou de l'arbre. C'est un archétype universel, une représentation qui vise autre chose que sa représentation.

Il n'est pas certain que de tels archétypes existent au sein des approches bottom-up ; et même si c'était le cas, il me semble prometteur de s'interroger sérieusement à ce propos pour mieux cerner le phénomène. Les régressions statistiques opérées par les réseaux de neurones ne pensent qu'à partir des

18. LESLIE G. VALIANT, *Knowledge Infusion : In Pursuit of Robustness in Artificial Intelligence*, Foundations of Software Technology and Theoretical Computer Science (Bangalore) 2008 (conférence), janvier 2008.

19. Cette citation est traduite de l'anglais par mes soins.

données elles-mêmes. Tout ce qu'elles ont sous les yeux, ce sont des cas connus. Elles peuvent certes produire de nouveaux exemplaires à volonté, mais il n'est pas sûr que ce processus de *génération* aveugle, qui engendre des représentations similaires à partir de représentations existantes, soit comparable au processus de *généralisation* — qui cherche à comprendre la structure sous-jacente à tous les cas représentables.

En termes techniques, je dirais que les catégories du réseau de neurones sont nominalistes : des mots tels que « arbre » ou « triangle » sont des raccourcis de langage commodes pour désigner d'un seul coup l'ensemble des arbres et des triangles représentables. Au lieu de quoi, l'esprit humain se réfère à des universaux : il ne parle pas seulement de tous les arbres connus ou imaginables, mais de l'arbre en général. Il n'est bien sûr pas certain qu'un « arbre en général » existe — c'est un débat qui remonte aux scolastiques du moyen-âge. Nous sommes néanmoins capables de penser le concept de « l'arbre en général ». Cet archétype est même si puissamment enraciné dans les structures de notre pensée qu'il est difficile de comprendre ce qu'est *ce chêne-ci* ou cet *érable-là* sans se référer implicitement à l'idée d'un « arbre en général ». C'est pourquoi l'image mentale, dans la cognition humaine, sert de liant entre le niveau bottom-up (le Système 1) et le niveau top-down (le Système 2). C'est une image qui s'appuie sur le monde sensible de ses deux pieds pour toucher des mains le monde des idées universelles.

Je vais consacrer les derniers chapitres de ma thèse à justifier la présence de ce troisième système cognitif, proprement intuitif, en tenant compte des paradigmes de programmation contemporains. Mais avant de m'engager dans cette voie épineuse, j'aimerais tout d'abord clôturer la critique des réseaux de neurones que je m'étais engagée à produire, en épinglant quelques différences clefs existant entre nos réseaux de neurones artificiels et le fonctionnement connu du cerveau. Ce point est important pour mon argumentaire, car il signifie que l'esprit humain ne se réduit pas forcément aux seules opérations décrites par les techniques de l'intelligence artificielle. S'il existe en effet un Système 3, ce système doit mobiliser des fonctionnements cérébraux différents de ceux que nous tentons déjà d'imiter avec nos réseaux de neurones.

7.4.2 Le Réseau de neurones vs le cerveau biologique

Le terme « réseau de neurones » peut laisser à penser, comme cela se produit trop souvent dans la presse vulgarisée, que les techniques d'intelligence artificielle bottom-up fonctionnent fondamentalement d'après les mêmes principes que le cerveau humain. Or il existe au moins cinq particularités du réseau de neurones qui le distinguent de notre organe cérébral : l'épuration, la digitalisation, la connectivité, la discrétisation, la localisation et une faible intégration. Je vais analyser ces cinq aspects les uns après les autres, avant de conclure sur une critique plus générale.

La première différence, que je qualifie *d'épuration*, vient de ce que le neurone formel de McCulloch et Pitts que j'ai décrit au chapitre 7 ne garde qu'une ossature très simplifiée du neurone biologique. Dans le modèle mathématique, on fait l'économie des neurotransmetteurs, des phénomènes physicochimiques ou du rôle joué par les autres cellules nerveuses (cellules gliales, cellules de Schwann, cellules de Purkinje, astrocytes, épendymocytes...). Cette simplification est, je le concède, inhérente à tout modèle mathématique. Il est toutefois douteux que le but premier du neurone artificiel soit de modéliser le neurone pour mieux en comprendre son fonctionnement biologique. Il s'agit plutôt de s'inspirer du neurone biologique pour créer un outil d'intelligence artificielle efficace. On souhaite en particulier que l'entraînement du réseau de neurones puisse être fait de façon *locale*, en améliorant itérativement les réponses du système à chaque nouvelle donnée disponible plutôt qu'en parcourant l'ensemble des données à chaque recalibrage. En ce sens, le réseau de neurones semble plus proche de la nature : il ne produit pas des réponses optimales, qui sortent toutes faites de la cuisse de Jupiter, mais il s'améliore de façon progressive, flexible et autonome — au gré des expériences.

La seconde différence, tout aussi évidente, entre le neurone formel et le neurone biologique, est ce que j'appelle la *digitalisation*. Le réseau de neurones artificiels ne traite pas des signaux physicochimiques, mais des nombres. Plutôt que de voir l'information comme une onde en déplacement au sein des cellules, les entrées et les sorties du neurone formel sont des symboles : il s'agira tantôt d'un nombre correspondant à l'intensité d'un pixel, tantôt d'une lettre de l'alphabet, tantôt du nombre 0 ou 1. En d'autres termes, tout se passe comme si le neurone ne faisait rien d'autre que manipuler de l'information abstraite, et non pas des potentiels électriques, des protéines, des cellules.

Une troisième différence concerne la *connectivité* du système. Les réseaux de neurones artificiels, on l'a vu, présentent des couches bien alignées. Les neurones d'une couche prennent en entrée les informations des neurones de la couche précédente et ils communiquent leurs réponses aux neurones de la couche suivante. Cette structure simplifiée optimise les méthodes statistiques implémentées par ces neurones artificiels. Elle ne correspond cependant pas du tout à la nature du cerveau humain. Dans ce dernier, il existe tout d'abord de nombreux types de neurones différents. Certains sont connectés uniquement de proches en proches, d'autres présentent des connexions traversant l'ensemble du système nerveux. Les neurones pyramidaux de l'hippocampe, par exemple, ont des dendrites connectées au noyau accumbens, à la formation réticulée du tronc cérébral ou à l'aire préoptique médiane, située dans l'hypothalamus.

Le quatrième point est ce que j'appelle la *discrétisation*. Les problèmes traités par les réseaux de neurones prennent un nombre entier de données en entrée (par exemple une grille de pixels) et fournissent une réponse au

sein d'un nombre entier de solutions possibles (« c'est un carré », « c' est un triangle », « c'est un cercle »...). Il est en outre commode d'assigner un neurone à chaque entrée et un neurone à chaque réponse possible. L'analyse d'une image se présente dès lors de la façon suivante : l'image est tout d'abord enregistrée par un capteur numérique, sous forme d'une grille de pixels. La valeur de chaque pixel est ensuite introduite en entrée à un neurone de la couche initiale du réseau de neurones, de sorte que chaque neurone représente exactement un pixel. On observe ensuite une succession de calculs statistiques, à l'issue desquels un neurone final s'active. Ce dernier correspond à une réponse possible, comme : « ceci est un triangle ! »

Si le cerveau fonctionnait de la même manière, l'on devrait s'attendre à ce que l'image observée par le sujet humain arrive au contact de sa cornée, que la cornée soit elle-même tapissée de récepteurs discrets, que ces récepteurs encodent chacun un pixel de l'image incidente et transmettent l'information à un neurone assigné. Il faudrait ensuite que ces neurones interagissent pour activer un seul autre neurone, qui correspondrait à l'objet perçu (triangle, carré, cercle...). Il faudrait qu'il y ait des neurones récepteurs, assignés deux à deux avec les capteurs de la cornée, mais également des neurones terminaux correspondant aux concepts que nous manipulons (triangles, carrés, cercles).

Or on peut contester que les images qui arrivent sur la cornée soient traitées par le cerveau de façon discrète, en découpant cette image en un ensemble de pixels. Il est important de noter que ce n'est pas parce que les cellules de la cornée sont des petites unités, de types cônes ou bâtonnets, qui reçoivent la lumière incidente pour donner jour à la vision, que cette vision est elle-même découpée en petites unités. Il se peut en effet que ces cônes et ces bâtonnets transforment une onde lumineuse incidente en une onde électrochimique. Cette onde est un processus continu, et non discret. Elle excite simultanément tous les capteurs de la cornée, quoique de façon différenciée. Ces capteurs, à leur tour, à l'instar d'une corde de guitare, émettent un signal qui se propage dans tout le tissu cérébral. Les neurones, à ce titre, servent de points de relais. Ils peuvent modifier l'onde cérébrale, l'atténuer, la renforcer. La propagation des informations dans le cerveau s'apparente peut-être à ce qu'il se passe lorsqu'une onde, à la surface d'un lac, heurte une pierre dépassant de la surface ? L'onde est déformée par la présence de la pierre (voir figure ). La déformation ne survient cependant que si la pierre dépasse la surface de l'eau. Dans le cas contraire, il n'y a pas d'interférence. De la même manière, leurs neurones pourraient créer des schémas d'interférences dans la propagation de l'onde cérébrale qui dépendraient du fait qu'ils soient ou non activés.

Cette métaphore que je propose ne dit rien d'autre que ceci : plutôt que de représenter le traitement cérébral uniquement de façon discrète, en voyant les neurones comme des unités indépendantes qui transportent chacun un petit quanta d'information, il est envisageable de penser que l'on puisse également, au moins hypothétiquement, envisager l'information du



FIGURE 7.15 – Des animaux se posant sur la surface de l'eau créent des ondes. Celles-ci sont déformées par la présence d'obstacles.

cerveau sous la forme d'une onde. Cette représentation continue du cerveau ne devrait pas, idéalement, entrer en contradiction avec la représentation discrète. Il s'agirait plutôt d'un modèle complémentaire. Au même titre que la physique quantique peut être décrite sous deux visages, en assimilant la lumière tantôt à des corpuscules, tantôt à une onde, je plaide en faveur de l'idée que le cerveau puisse aussi être vu tantôt comme un processus digital et tantôt comme un processus analogique. Dans un cas comme dans l'autre, on ne modélise pas les mêmes aspects du phénomène. On n'obtient donc pas nécessairement la même compréhension de ce qu'est la cognition humaine.

La dernière différence existant entre les neurones artificiels et biologiques, que j'appelle la *localisation*, devrait donner encore un peu plus d'ampleur à mon argumentaire. Dans le réseau de neurones artificiels, tout se passe comme si chaque neurone avait un rôle assigné de toute nécessité. Pour chaque processus donné (vision, ouïe, raisonnement), il existe des neurones initiaux qui encodent l'information et des neurones terminaux qui apportent une réponse. Or dans le cerveau, on observe une grande plasticité. Certaines régions neuronales peuvent intervenir dans différents rôles (vision, régulation, compréhension des émotions...). De plus, le rôle de ces régions peut changer au cours du temps. L'exemple le plus frappant a été observé en 2006 lorsqu'un patient, à la suite d'un IRM, a découvert qu'il avait perdu 90% de sa matière cérébrale. Il était pourtant tout à fait capable de se mouvoir, de se déplacer et de vivre normalement. De manière générale, l'ablation d'une partie du cerveau conduit les aires corticales survivantes à assurer les fonctions neuronales qui étaient auparavant remplies, assurées par les cellules manquantes. On pourrait même se demander si certaines tâches normalement allouées au cerveau ne puissent pas être prises en charge, dans une certaine mesure, par d'autres organes du corps humain.

Ce phénomène de plasticité est connu depuis les travaux de Jean-Marie Flourens (1794-1897). Ce dernier procédait à la résection de certaines parties du cerveau du pigeon et observait que les volatiles étaient capables, après l'opération, de se rééduquer rapidement. Ce qui semble prééminent,

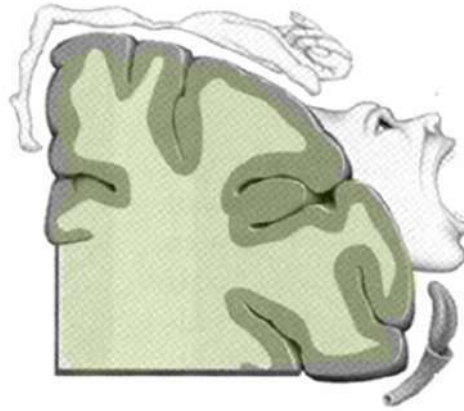


FIGURE 7.16 – Dans le cortex sensorimoteur, chaque zone cérébrale correspond à une zone du corps. Cela ne veut pas dire que la correspondance se fait point par point, mais elle est très nette. Certaines parties du corps, comme les mains, la langue et le visage, sont surreprésentées. Le volume du tissu nerveux associé à ces organes occupe la majeure partie du cortex somesthésique.

sur le plan comportemental, n'est pas tant la localisation de la résection, que la quantité de matière ôtée. Cela suggère de façon magistrale l'idée d'une certaine plasticité cérébrale, sans signer toutefois l'acte de décès de toute forme de localisation. Et pour cause : en marge de cette plasticité, et simultanément à celle-ci, il existe *aussi* des cas spectaculaires pour lesquels les neurones possèdent en effet des rôles précis. On peut penser notamment à l'homoncule de Penfield, dans le cortex somesthésique, où chaque zone corticale s'active sélectivement lorsque nous sentons une partie de notre corps : certaines neurones sont liés aux doigts, aux lèvres, aux oreilles ou aux genoux (voir figure 7.16).

Je pense qu'il est raisonnable de penser que si les réseaux de neurones capturent en effet une certaine dimension du fonctionnement du cerveau humain, il serait naïf, pour ne pas dire tout à fait téméraire, d'ignorer les spécificités de ce dernier. La grande richesse du cerveau vient de sa haute intégration. Les cellules nerveuses sont hautement connectées, non seulement entre elles, mais également à l'ensemble de l'organisme, si bien que l'activité du cerveau est corrélée avec celle du cœur, des intestins, des muscles, etc.

Les réseaux de neurones réduisent cette complexité pour faire correspondre, à chaque neurone, une information discrète précise. Il s'ensuit que tout l'art de l'intelligence artificielle, aujourd'hui, consiste à recomposer *a posteriori* les liens existant entre les neurones, de façon à obtenir une compréhension globale de l'information dont ils sont porteurs chacun individuellement. C'est ainsi que l'analyse d'images, par exemple, nécessite de lier

chaque pixel les uns les autres, en utilisant des poids synaptiques, pour comprendre si ces pixels, pris ensemble, constituent ou non une forme connue. Autrement dit, chaque neurone formel est un point isolé et les synapses existant entre ces neurones sont des questions binaires consistant à se dire : ces deux points sont-ils liés d'une certaine manière ?

À l'inverse de cette approche *discrète* et *localisée*, il se pourrait que le cerveau humain reçoive une information globale, sous la forme d'une onde qui stimule toutes les cellules du corps en simultané, quoique de façon hautement différenciée. Qu'est-ce que je veux dire par là ? Que l'onde se propage, non pas *uniquement* d'une cellule à l'autre par saut de puces, mais *aussi* à l'ensemble des cellules comme une grande vague. Comme l'onde n'est pas de forme égale partout, des différences de potentiel électrique apparaissent dans certaines régions du cortex. C'est ce contraste de l'activité électrochimique des neurones, plus encore que l'intensité de l'onde elle-même, qui permettrait au cerveau d'identifier des formes connues. La vision d'une pomme, par exemple, créerait des différences d'activation entre les cellules du cerveau qui ne seraient pas comparables aux différences d'activation que le sujet éprouve en présence d'un arbre.

L'idée que l'information du cerveau est issue d'une activité différenciée des cellules fait écho, de façon certes analogique, à la théorie de l'information intégrée de Tononi. Dans cette théorie, les processus cognitifs tels que la sensation, la perception ou la conscience ne sont pas assignés à des zones précises du cerveau, mais à la capacité des cellules nerveuses à entrer en résonance et à se différencier statistiquement, par leur activité synchronisée, de l'activité des autres cellules du cerveau^[20].

Pour conclure, il se pourrait que l'esprit humain fonctionne de façon holistique, en partant de l'information globale de la perception et en la décomposant en parties indépendantes correspondant aux objets, aux points de couleurs, correspondant chacune à des schémas d'activation différenciée bien spécifiques du tissu cérébral. En lisant cette conclusion, le lecteur ne devrait pas être surpris outre mesure. Je n'ai fait ici que transposer la critique kantienne du chapitre 3 au fonctionnement plus spécifique du cerveau et des réseaux de neurones. Dans cette hypothèse que je fais, les algorithmes contemporains fonctionnent de façon non pas holistique, mais séquentielle : ils additionnent les unes aux autres des informations isolées, représentées chacune par un neurone spécifique, dans l'espoir de se saisir des rapports

20. L'explication que je donne ici est une schématisation simplifiée. D'un point de vue plus technique, Tononi et ses collègues définissent une mesure d'information, inspirée des travaux de Shannon, qu'ils appellent l'information intégrée. Cette grandeur cherche à mesurer si un groupe de cellules nerveuses est statistiquement indépendant des autres cellules du tissu cérébral. Cela veut dire qu'on cherche des schémas d'activation, des formes, au sein du tout que constitue l'activité organique du cerveau. Voir notamment TONONI G. et al., *Integrated information theory (IIT) 4.0 : Formulating the properties of phenomenal existence in physical terms*, arXiv, 30 décembre 2022.

que ces points d'information entretiennent les uns les autres. Dans la dernière partie de ma thèse, je souhaiterais maintenant rassembler toutes les briques critiques que j'ai abordées jusqu'à présent de façon à reconstituer une théorie tricamérale de l'esprit humain. Cette théorie devrait justifier la présence d'un Système 3 entre les Systèmes 1 et 2 identifiés par les travaux de Daniel Kahneman. Elle devrait en outre montrer clairement en quoi constitue l'angle mort de l'intelligence artificielle.

— Ce qu'il faut retenir de ce chapitre —

La deuxième stratégie de programmation à avoir vu le jour est la stratégie *bottom-up*. Contrairement à l'approche top-down, cette stratégie ne cherche pas à doter la machine de règles de résolution générales, mais à utiliser les données disponibles sur le problème pour permettre à cette machine d'apprendre par elle-même.

L'approche bottom-up la plus populaire est le réseau de neurones. Ce modèle, inspiré du neurone biologique, est formalisé dans les années cinquante par McCulloch et Pitts. Il se construit suivant deux stratégies d'apprentissage :

- Dans *l'apprentissage supervisé*, les concepts rationnels sont définis par le programmeur. Les utilisateurs donnent un feedback à l'algorithme chaque fois que ce dernier commet une erreur dans la manipulation de ces concepts.
- Dans *l'apprentissage non supervisé*, la machine crée ses propres catégories, mais celles-ci ne sont pas dotées d'une définition rationnelle ; il s'agit du simple rassemblement statistique d'objets présentant, entre eux, des similarités.

En dépit de ces deux stratégies, les réseaux de neurones peinent à abstraire les propriétés générales des objets qui lui sont présentés. Cette lacune suggère qu'il existe, chez l'être humain, un troisième niveau de pensée entre le niveau top-down et le niveau bottom-up. C'est cette hypothèse qui sera envisagée dans le prochain chapitre.

Chapitre 8

L'Angle mort de l'IA : Esquisse d'une théorie holistique du jugement

J'aimerais résumer ce qui a été dit jusqu'à présent à l'aide d'une illustration issue du jeu d'échecs. Cela va me demander de faire une brève digression. Depuis la victoire de Deep Blue sur Kasparov, en 1996, très peu d'êtres humains se sont risqués à affronter la machine dans des rencontres officielles. Il n'a pas fallu longtemps, en effet, après cette première percée, pour que l'intelligence artificielle gravis triomphalement les marches qui la séparaient encore de l'invincibilité. En quelques années seulement, des logiciels tels que StockFish, Rybka ou Deep Fritz étaient si performants que les rares têtes brûlées qui acceptaient de les affronter se sont rendues victimes d'une capitulation^[1]. La machine s'est peu à peu imposée comme moteur d'analyse, arbitre et coach, tant sa maîtrise du jeu est devenue profonde. Les chances de victoire entre le champion du monde en titre et StockFish sont estimées statistiquement à un pour cent^[2]. Des joueurs d'échecs continuent néanmoins d'enregistrer des victoires ça et là, en procédant par essais-erreurs. Ils s'appuient pour ce faire sur une limitation que l'on appelle l'effet d'horizon (voir figure 8.1).

1. Kramniz affronte une première fois Deep Fritz 7 en 2022 et obtient finalement, après un but début (3-1), un match nul (4-4). Lors de la seconde rencontre en 2006, peu après avoir obtenu le titre de champion du monde, il affronte Deep Fritz 10 et finit avec 4 nulles, 2 défaites et zéro victoire. Ce match lui promettait un million de dollars en cas de victoire et la moitié de la somme en cas de défaite. en dépit de ces gains généreux, plus aucune rencontre officielle n'a eu lieu entre la machine et un champion du monde.

2. Ce calcul est réalisé en comparant le score ELO maximum de Magnus Carlsen sur toute sa carrière (2 889) et de Stock Fish à la date de mai 2021 (3 548). La différence de 659 points suggère, d'après l'échelle logistique utilisée pour calculer le score ELO, que les chances de victoire de Magnus sont très exactement de 1%. Il est toutefois généralement admis qu'une différence de 600 points à haut niveau de compétition a plus d'importance que ne le suggère le simple écart de points.

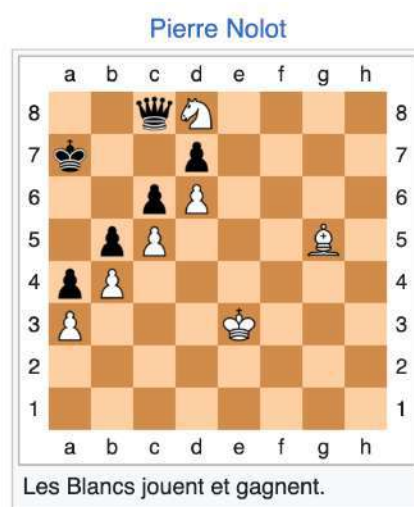


FIGURE 8.1 – **Effet d’Horizon** : Cette position est soluble par un humain mais complexe pour une machine. Les Blancs gagnent en amenant le roi en e7, puis le fou en f8, puis en déplaçant le cavalier pour prendre le pion noir en d7. Il faut ensuite pousser la promotion du pion en d6.

L’effet d’horizon est lié à la difficulté, pour une intelligence artificielle, d’établir des stratégies à long terme en l’absence de toute espèce d’avantage à court terme³. L’être humain, face à la complexité calculatoire d’une position échiquéenne, a tendance à se rabattre sur des concepts synthétiques : le contrôle des diagonales, la structure de pions, etc. Ces concepts, acquis par un mélange d’expérience et de raisonnement, sont suffisamment clairs pour être manipulés intuitivement sans devoir calculer toutes les possibilités ; mais ces concepts restent trop complexes, ou trop vagues, pour être exprimés de façon quantitative dans le langage de la machine. Quelle valeur faut-il donner au contrôle d’une colonne ? Quelle valeur donner à la prise d’une pièce ? Comment définir ce qu’est une bonne structure de pions ?

Il va de soi que si la machine pouvait envisager toutes les possibilités, depuis le coup d’ouverture jusqu’au dernier coup de la partie, elle n’aurait pas besoin de manipuler ces concepts simplificateurs. Elle aurait seulement sous les yeux, dans toute son extension, l’arbre des possibilités⁴. Le jeu serait limpide pour elle. Or dans la pratique, il est impossible pour une

3. La définition exacte de l’effet d’horizon demande un peu plus de vocabulaire technique. Il se produit lorsque la machine, après avoir anticipé n coups, estime avoir trouvé un bon mouvement alors que, si elle regardait le coup $n + 1$, elle se rendrait compte que ce n’est pas le cas. Inversement, elle peut estimer qu’une situation est mauvaise après n coups alors qu’en allant plus loin elle aurait une toute autre opinion.

4. Cette appellation "arbre des possibilités" vient de ce que chaque coup peut être représenté au sein d’un arbre. La base de cet arbre correspond au coup d’ouverture et les ramifications sont tout autant de réponses possibles.

intelligence artificielle d'envisager tous les mouvements, car ceux-ci croissent de façon exponentielle. En supposant qu'il n'existe que 20 coups possibles lors de chaque tour⁵ et qu'une partie ne fasse pas plus de 60 coups⁶, on enregistre 20^{60} possibilités, ce qui est très proche du nombre estimé d'atomes dans l'univers. Cela dépasse clairement la capacité des machines connues aujourd'hui.

Dans la mesure où l'arbre des possibilités, aux échecs, est bien trop grand pour être exploré dans son entier, la machine doit, comme l'être humain, éliminer *a priori* les coups qui semblent les moins avantageux. Cela lui demande d'évaluer une position à l'aide de concepts : valeur des pions, contrôle des colonnes, structure de pions, etc. Tout le problème de l'informatique consiste donc, dans sa substantifique moelle, à doter la machine d'une capacité à réunir sous une même expression des situations de jeu pourtant différentes, en créant au moins trois catégories générales : les « situations gagnantes », les « situations nulles » et les « situations perdantes ». Une situation est dite gagnante lorsqu'elle présente de meilleures chances de victoire que celle qui est observée actuellement ; elle est dite perdante lorsque les chances de victoire diminuent. Si ces catégories sont construites de façon intelligente, alors l'algorithme, plutôt que d'analyser bêtement toutes les possibilités, peut arrêter son analyse dès qu'un avantage suffisamment grand ou suffisamment faible est observé. On appelle cela le *pruning*.

J'ai suggéré, dans les chapitres précédents, que ces catégories peuvent être créées de deux manières. L'approche top-down consiste à doter la machine d'un ensemble de règles analytiques, appelées généralement, dans le jargon informatique, des « heuristiques ». Ces règles assignent une valeur à chaque situation de jeu en regardant uniquement la position des pièces à un instant donné ; elles peuvent donc s'exprimer sous forme de lois logiques quantifiables, consistant à regarder chaque case, chaque pièce, et en vérifiant si des menaces sont exercées. L'exemple le plus célèbre est formulé par Alan Turing, lorsqu'il suggère d'assigner une valeur à chacune des pièces : 10 pour la dame, 5 pour une tour, 3 pour un fou, 2,5 pour un cavalier et 1 pour un pion⁷. Cette règle du pouce, bien connue des débutants, est cependant très souvent prise en défaut. En sacrifiant une pièce réputée plus forte, un

5. Il existe seize coups à l'ouverture : déplacer un pion de deux cases, déplacer un pion d'une seule case ou déplacer un cavalier. Ce nombre s'accroît rapidement peu après l'ouverture et diminue en fin de partie.

6. Certaines parties enregistrées ont duré plus de 300 coups. Or il n'y a pas de raison pour que ce nombre augmente davantage. Une partie ne prend fin, d'après les règles officielles, que lorsque les joueurs répètent trois fois la configuration de jeu. Comme le nombre de ces configurations est limité, une partie finit toujours par s'arrêter, mais cela peut prendre énormément de temps. C'est pourquoi la machine doit être dotée de concepts synthétiques, lui permettant d'évaluer lorsqu'une position est bloquée ou tourne en rond, plutôt que de procéder à la force brute.

7. À noter que cette règle connaît quelques variantes. On donne parfois 9 points à la dame au lieu de 10, et il arrive aussi de donner 3 points aux cavaliers au lieu de 2,5.

joueur adroit pourra se créer un avantage positionnel. Typiquement, il se peut que le sacrifice de la reine conduise à un échec et mat, ou que le sacrifice d'un fou compromette tellement la défense du roi de l'adversaire que ce dernier sera, à son tour, obligé de sacrifier du matériel. C'est pourquoi on comprendra que la règle du pouce, seule, ne suffit pas à décrire la valeur d'une position. La question fondamentale qui s'impose est la suivante : combien de règles faut-il pour avoir une description satisfaisante ? Et plus fondamentalement : est-il possible de ramener toute la complexité du jeu à de pures règles analytiques ?

L'approche bottom-up, quant à elle, prend un tout autre parti. Plutôt que de donner des règles d'analyse prêtes à l'emploi, elle préfère se baser sur les statistiques. Elle commencera par se doter d'un grand répertoire de parties jugées « solides », jouées préférentiellement par les joueurs du plus haut niveau. Lorsqu'elle doit évaluer une position donnée, elle regarde dans la base de données quel pourcentage de parties ayant eu cette position a été gagné par les Noirs et quel pourcentage a été gagné par les Blancs. Cela lui donne une vision synthétique, purement empirique, du jeu. Bien sûr, il peut arriver que la situation à évaluer ne figure pas dans sa base de données, dans quel cas elle doit se rabattre sur des règles analytiques. Ces règles sont utilisées jusqu'à ce que la machine retrouve une situation connue. Si elle n'aboutit pas à une situation connue, elle s'arrête à une profondeur donnée (la profondeur est le nombre de coups que la machine anticipe).

On comprend donc que les règles analytiques et les règles synthétiques marchent main dans la main. L'approche bottom-up, lorsqu'elle s'appuie sur des techniques d'auto-apprentissage, ne va cependant pas encoder par avance les règles analytiques ; elle va plutôt donner à la machine des méthodes permettant d'extraire des règles efficaces à partir de sa propre base de données. L'IA utilise donc un certain sous-ensemble des parties répertoriées dans sa base de données pour déduire des règles d'analyse efficaces ; et elle utilise les autres parties répertoriées de sa base de données pour tester l'efficacité de ces règles. Une fois que les règles se sont solidifiées, la machine peut ensuite entreprendre des parties contre elle-même, un peu à la façon de la méthode évolutionnaire, pour créer de nouvelles bases de données ou pour renforcer encore l'efficacité de ses heuristiques. Dans le cas extrême, la machine peut apprendre à jouer aux échecs en commençant sans aucune autre information préalable que les règles du jeu. C'est le cas du programme Alpha Go, qui apprend sans base de données préalable⁸. Cette méthode d'apprentissage est appelée l'apprentissage non supervisé, ou encore, l'apprentissage par renforcement.

En dépit de l'évolution incontestable des machines sur l'échiquier, l'in-

8. Alpha Go est rigoureusement parlant le nom donné au programme de Go ayant vaincu Lee Sedol en 2017. Leela Chess Zero, ou Lc0, est le nom donné au logiciel d'échecs s'inspirant de la méthodologie adoptée par Alpha Go.



FIGURE 8.2 – Cette figure illustre la stratégie adoptée par Hikaru Nakamura (2800 points ELO) pour défaire le programme Stockfish (3500 ELO) sur l'échiquier.

telligence artificielle ne s'est pas totalement émancipée de l'effet d'horizon, si bien que des joueurs expérimentés parviennent encore à vaincre la machine çà et là. Un exemple patent de ce phénomène est démontré par Hikaru Nakamura, classé deuxième meilleur du monde, dans un match contre Stockfish du niveau 7 (3200 points ELO) à la date du 21 mars 2024⁹. Au lieu d'affronter la machine dans un match régulier, il décide de mener des sacrifices désavantageux de façon à créer une structure de pions fermée (voir figure 8.2), où les pièces n'ont plus aucune possibilité de jeu. Hikaru sait néanmoins qu'un cavalier, bien que supposé plus faible d'un fou d'après la règle du pouce, s'avérera plus puissant dans ce genre de configuration. Après 271 mouvements, composés en grande partie de va-et-vient, il finit par faire promouvoir six pions de l'autre bout de l'échiquier — qu'il transforme en fous de façon à ajouter, au mérite de son écrasante victoire, le panache d'un échec et mat artistique.

Tout se passe — c'est ce que je me suis efforcé de montrer dans les chapitres précédents — comme s'il existait deux registres de jugements. Le jugement analytique consiste à prendre une décision sur base de concepts rationnels et calculables, constitués de règles quantitatives. Le jugement synthétique, quant à lui, repose sur l'expérience, les statistiques, la mémoire. Les joueurs humains comme les machines ont recours à ces deux composantes face à l'échiquier. J'aimerais toutefois envisager l'idée qu'il existe, au sein de l'esprit humain, des concepts qui occupent l'exact milieu entre un jugement synthétique pur et un jugement analytique pur. Ces concepts seraient, dans la théorie kantienne, l'analogue du jugement synthétique *a priori*, quoiqu'il ne soient pas forcément définis de la même manière. Ce sont des représentations qui peuvent faire l'objet de calculs, mais sans pouvoir être pleinement

9. Notons qu'à la date de la rencontre, StockFish est considéré comme le meilleur programme d'échecs du monde. Il a remplacé certaines de ces heuristiques naïves, qui lui avaient valu des défaites contre AlphaGo en 2017, par un réseau de neurones robuste.

et totalement explicitées. Ce sont des façons de réduire la complexité d'une situation de jeu à quelques schémas simples, manipulables, qui ne sont toutefois pas réductibles (1) ni à un ensemble de règles rationnelles quantitatives, ni (2) à un ensemble de statistiques empiriques mémorisables.

Pour conclure, j'espère être parvenu à suggérer, en m'appuyant sur l'exemple du jeu d'échecs, qu'il existe trois niveaux de pensée : (i) le niveau synthétique, correspondant à des associations statistiques (un répertoire de partie est associé à une probabilité de victoire), (ii) le niveau analytique correspondant à des inférences logiques (des calculs sont faits pour évaluer une position, en utilisant par exemple la règle du pouce) et (iii) un niveau intuitif, qui résiste aussi bien à l'association statistique qu'à des règles logiques et quantitatives. J'aimerais consacrer le prochain chapitre à justifier l'existence de ce niveau intermédiaire et à en montrer la pertinence. Je reprendrai pour ce faire l'exemple des échecs. Un autre chapitre sera ensuite dédié à comparer ces trois niveaux et leurs relations respectives.

8.1 Les Concepts intuitifs

Reprenons quelques instants l'exemple du jeu d'échecs. Contrairement à la machine, le joueur humain qui estime qu'une position d'échecs est bonne en raison du contrôle d'une colonne, par exemple, ne peut pas énumérer explicitement, quand bien même il le voudrait, l'ensemble des parties où une telle position lui a valu la victoire. Il ne fait que reproduire un schéma qui s'est avéré juste ; ce schéma est une idée simplificatrice, ou encore, une abstraction, qui consiste à isoler une composante (la colonne ouverte) de l'ensemble de la partie. Lorsqu'il y recourt, il fait comme si la colonne était indépendante des autres facteurs intervenant dans le problème. Il lui donne, pour ainsi dire, une vie autonome. Il sait pourtant qu'aucune partie ne se réduit à ce simple concept de « colonne ouverte ». Il est conscient de l'écart existant entre son concept et l'objet auquel il se rapporte. Cette différenciation entre les mots et les choses, entre les propriétés et les objets, est la marque indélébile des concepts intuitifs.

De tels concepts existent-ils dans les algorithmes actuels ? On pourra objecter que le réseau de neurones ne fait pas autre chose lorsqu'il calibre les poids de son réseau. Il isole certains paramètres de la masse des données qui lui ont été soumises et adapte par conséquent les relations existantes entre les neurones. De cette façon, ses réponses à certains problèmes tiennent compte des exemples auxquels il a été confronté lors de son apprentissage, quand bien même ces exemples seraient, ultérieurement, ôtés de sa mémoire. Pour le dire autrement, l'évolution structurelle du réseau de neurones semble être capable de s'adapter aux cas qui lui sont présentés sans qu'il lui soit nécessaire de se rappeler explicitement chacun de ces cas.

Pour la machine, il est important, au moment de construire son réseau de

neurones, que l'information lui soit présentée comme un ensemble de symboles : typiquement des nombres binaires. Le réseau de neurones traite ces symboles en les parcourant un à un. Cela suppose qu'il existe, pour ainsi dire, des unités minimum dans la structure de la pensée, des éléments informationnels atomiques. Lorsqu'il s'agit d'analyser une image, par exemple, ces unités sont les pixels. Pour permettre à la machine d'extraire des catégories au sein de cette image, comme la catégorie de « colonne ouverte » au sein d'une position échiquienne, il suffit de présenter l'information comme un tableau de nombres. Chaque nombre du tableau représente la valeur du pixel (ce nombre représente la luminosité du pixel dans le cas d'une image en noir et blanc. Dans le cas d'une image en couleurs, il faut trois nombres reprenant la teneur en bleu, en rouge et en vert). Or, il n'est pas garanti que l'être humain, dans le cas spécifique où il cherche à abstraire une catégorie au sein d'une image, réfléchisse à partir des pixels, et ce pour deux raisons : tout d'abord, cela suppose que l'esprit humain découpe le tout de la perception en unités minimales ; ensuite, il faut admettre qu'il soit capable d'énumérer une à une ces unités pour les traiter et en extraire des conclusions.

Ces deux suppositions se heurtent à un problème épistémologique : l'être humain peut-il identifier des pixels sans avoir préalablement un concept de couleur ? Est-ce la couleur qui est définie à l'aide de pixels ou sont-ce les pixels qui sont définis à l'aide de la couleur ? Dans la machine, on ne se soucie pas de cette interdépendance existant entre la couleur et les pixels. On assigne des nombres et on suppose que l'ordinateur chargé de traiter ces nombres dispose de règles de manipulation concernant l'addition, la multiplication, etc. Dans le cas de l'être humain, cependant, et à moins de supposer que l'esprit soit dans le cerveau et que le cerveau soit lui-même une machine à calculer — mais ceci rendrait l'argument circulaire — on observe une interférence entre le tout et les parties. L'identification d'un point de couleur à une position donnée est influencée par les points de couleur adjacents, comme cela est manifeste par l'illusion de Valois (voir schéma). Même dans des images en noir et blanc, la vision des contrastes est influencée par le background, comme dans l'illusion de Chubb (voir schéma).

Ces illusions suggèrent que l'esprit humain ne raisonne pas à partir de points prédéfinis, mais à partir des contrastes. Le découpage du tout de l'image a bel et bien lieu, mais il ne consiste pas à inscrire arbitrairement chaque point de l'image dans une grille de pixels préétablie. Il consiste bien plutôt à isoler des zones qui se différencient les unes des autres. Ce découpage est dynamique. Dès qu'une forme a été identifiée et qu'elle est abstraite de l'image comme une entité à part entière, il reste possible de décomposer, à son tour, cette forme en zones plus petites. De fil en aiguille, l'esprit humain peut arriver à des points, mais ceux-ci ne sont pas des unités minimales de la perception ; ce sont des points d'ancrage momentanés. C'est la raison pour laquelle il n'existe rien de tel, dans l'expérience concrète du peintre dans son atelier, qu'un pur point de couleur. Chaque point a une dimension, une

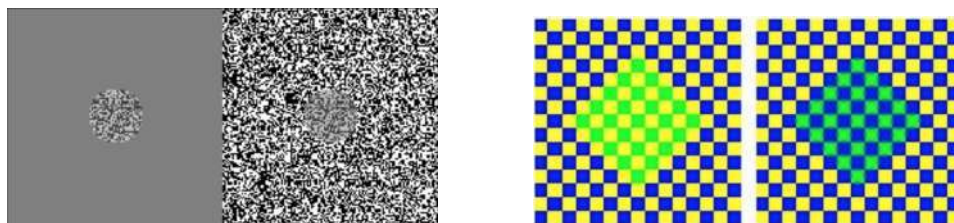


FIGURE 8.3 – Dans l’illusion de Chubb (à gauche), les deux disques sont identifiables, mais celui qui nous apparaît sur un fond gris homogène nous semble plus contrasté. Dans l’illusion de Valois (à droite), les carrés de couleur vert sont identiques dans les deux images, mais la présence de carrés bleu dans le fond va les rendre plus bleutés à notre regard. Ces deux expériences suggèrent que la perception des couleurs et des contrastes fonctionne en isolant des composantes par comparaison.

forme, et il existe en son sein parfois tout autant de contrastes qu’avec le reste de la peinture.

Ces remarques que j’ai faites suggèrent la chose suivante : l’esprit humain possède, en plus des jugements synthétiques et analytiques, une approche holistique. Celle-ci consiste à découper le réel en procédant à des différenciations. Par un jeu mouvant de contrastes, il est possible d’extraire des catégories abstraites à partir du tout de la perception. Ces catégories perceptives sont ce à travers quoi l’esprit humain peut appréhender le monde sensible. Sans elles, l’expérience serait inintelligible : ce ne serait pas un tableau fait de couleurs distinctes, que l’on peut regarder les unes après les autres, mais un tout indivisible. Une réalité donnée d’un seul bloc dans l’esprit, sans objets, sans propriétés.

Ces catégories perceptives sont donc nécessaires à la compréhension du monde, et jouent en ce sens un rôle analogue à celui que Kant conférait aux catégories de l’Entendement. Elles n’ont cependant pas d’existence indépendante du monde lui-même. Il est tout aussi impossible (1) de se représenter le monde sans le faire au travers de ces catégories que (2) de se représenter les catégories en dehors du monde. Le point (2) est la raison pour laquelle les catégories que j’ai en vue ici ne peuvent pas être définies au moyen de règles rationnelles et quantifiables. Ces catégories n’ont un sens que parce qu’elles sont connectées à un contexte empirique. Le point (1) implique quant à lui que la catégorie est autre chose que la simple énumération de faits sensoriels préalables. Elle est au contraire ce qui permet de penser les faits sensoriels et de leur donner une existence indépendante au sein de l’esprit humain.

C’est pourquoi les concepts intuitifs que le joueur d’échecs manipule lorsqu’il observe un plateau de jeu ne sont (1) pas la simple énumération des situations existantes et (2) qu’ils ne peuvent pas être définis à l’aide de règles formelles. Les concepts de « colonnes ouvertes », de « bonnes structures de

pions », « de contrôle » sont le résultat d'un découpage idiosyncratique du réel. Ce sont les points d'ancrage, les contrastes à partir desquels le joueur est capable de percevoir le plateau de jeu. Sans eux, ce plateau serait un tout sans intelligibilité. Y être confronté s'apparenterait sans doute à ce qu'éprouve le débutant lorsqu'il se sent perdu au milieu de tant de pièces, de cases, de menaces. Ce débutant, incapable de voir le plateau de jeu en termes purement échiquéens, se raccroche aux catégories qu'il connaît déjà : il s'arrête sur la forme géométrique des pièces, sur leur matériau, sur leur couleur, sur ce qu'elles évoquent en lui (un roi, un chevalier, une reine, un soldat...). Si on changeait la forme du jeu, que l'on remplaçait les authentiques pièces Staunton par des pièces modernes ou abstraites, si le plateau en bois devenait de la pierre, ou que le jeu était représenté à plat en deux dimensions, il serait perdu ; son analyse en souffrirait.

Le joueur d'échecs expérimenté, au contraire, analyse le jeu avec les catégories échiquéennes. Il décompose le plateau de jeu sans tenir compte de la texture, de la couleur, de la forme des pièces, mais en s'attardant uniquement sur les contrastes qui importent pour le jeu lui-même. Il ne voit plus des pièces isolées, dotées de formes géométriques ; il voit des structures de pièces interdépendantes, des liaisons, des symétries, des rapports de force. C'est pourquoi les concepts de « structures de pions » ou de « sécurité du roi » acquièrent tout à coup un sens intuitif à ses yeux. Ce n'est pas en raison d'une définition formelle, faite de règles et de calculs explicites, ou d'une anticipation des coups adverses. Ce n'est pas non plus parce qu'il se rappelle les parties antérieures ou parce qu'il a identifié des régularités (bien que cela intervienne aussi en grande partie). C'est parce que sa façon de décomposer le tout de la perception (le plateau de jeu dans son ensemble) en partie significative (les unités d'analyse) suit une autre structure. Pour le dire autrement, il possède une meilleure capacité d'abstraction intuitive.

Cette observation que j'ai posée explique que, lorsque l'on demande à deux joueurs d'échecs, un débutant et un maître, de mémoriser une configuration de jeu issue d'une vraie partie, le maître présente plus de facilité ; mais lorsque la configuration de jeu ne vient pas d'une vraie partie, mais d'une disposition aléatoire des pièces sur le plateau, c'est le débutant, au contraire, qui s'en sort le mieux. Et pour cause : sa vision du jeu n'est pas structurée par des concepts échiquéens, mais uniquement par des concepts géométriques. Le problème s'apparente, pour lui, à mettre des formes à certains emplacements. Le maître, au contraire, lorsqu'il voit une configuration de jeu défiant toutes les règles des échecs, éprouve un sentiment de désorientation similaire à celui du débutant qui découvrait le jeu pour la première fois. Tout lui semble bizarre, étrange, incongru. Il ne peut pas aisément découper sa perception en unités intelligibles.

Bien sûr, on pourra arguer que ces différences que j'ai décrites, entre le maître échiquéen et le débutant, sont analogues à ce qui se passe avec un réseau de neurones tel que AlphaGo. Au sein de ce dernier, des catégo-

ries sont construites au fur et à mesure que la machine est confrontée à de nouvelles situations de jeu, ce qui change la manière dont l'information est structurée. Cette vision est supportée par le fait que les réseaux de neurones sont aujourd'hui capables de vaincre n'importe quel joueur humain sur un échiquier. J'ai toutefois montré plus haut que l'effet d'horizon existe toujours, montrant que ces catégories de l'intelligence artificielle, quoique plus efficaces en moyenne que l'être humain, semblent répondre à des règles différentes. Elles se fondent sur des régularités statistiques entre les pièces, et non sur des contrastes globaux. Elles partent des parties et s'élèvent vers le général ; au lieu de quoi, l'on est en droit de se demander si l'esprit humain ne raisonne pas à partir du tout ?

Je reconnais que l'existence même d'une capacité de jugement holistique au sein de l'être humain reste purement hypothétique. Je m'estime satisfait, à l'issue de ce chapitre, si j'ai réussi à jeter un peu de lumière sur un angle mort des sciences informatiques. Il se pourrait, en dernière analyse, que les concepts intuitifs que j'ai décrits se ramènent à des statistiques. Je pense néanmoins qu'on a tort de le présupposer. Je voudrais en effet envisager la possibilité que les concepts intuitifs occupent le milieu entre une approche purement analytique, fondée sur la rationalité, et une approche purement statistique, fondée sur l'expérience. C'est pourquoi, dans le prochain chapitre, je vais plaider en faveur d'une représentation de l'esprit humain disposant de trois niveaux : le niveau top-down de la Rationalité, le niveau bottom-up de la Sensation et le niveau holistique de la Perception. Je vais tenter de montrer quel rapport cette théorie présente avec les théories de Kant et de Platon qu'on a vues dans les sections précédentes.

8.2 Une théorie tricamérale de la cognition

J'ai suggéré que les concepts intuitifs, s'ils représentent bien autre chose que de simples catégories statistiques, doivent occuper un milieu entre l'expérience et la rationalité. C'est pourquoi j'aimerais maintenant défendre que le jugement humain, tout comme la démocratie, est composé de plusieurs chambres parlementaires, qui exercent chacune un rôle différent dans la façon de traiter l'information et de délibérer.

La première chambre de l'esprit se contente de légiférer. Elle pose des règlements logiques et les analyse minutieusement. Il s'agit de la Rationalité. Les concepts qu'elle manipule sont rigoureusement définis. Ce sont des concepts abstraits et mathématiques, quantifiables et implémentables au sein d'une machine de calcul. Ses raisonnements sont toutefois fort éloignés des choses du monde.

La deuxième chambre est au plus près de l'expérience quotidienne. Elle prend des décisions semi-automatiques dans les affaires courantes de la vie. Il s'agit de la Sensation. Elle s'intéresse aux choses pratiques du monde. Elle

tisse des corrélations. Elle sait que la loi, ici, n'est pas définie de façon stricte, mais uniquement de façon statistique. Elle manipule non pas des règles bien définies, mais des règles implicites, pratiques et souvent inconscientes, que j'appelle les percepts. Ces derniers sont le fruit de l'expérience. Ils ne peuvent pas être décrits entièrement au moyen du langage.

La troisième chambre, pour finir, est celle des concepts intuitifs, qui n'ont pas de définition rigoureuse, mais sur lesquels on peut néanmoins porter des jugements. Ces jugements vrais sans justification interviennent lorsqu'un cas se présente qui ne peut pas être élucidé à l'aide de lois ni à l'aide de bonnes pratiques exécutives. Cette chambre que je décris est la Perception. Elle se saisit tantôt de la Rationalité, tantôt de la Sensation, s'efforçant de corriger leurs limitations. La Perception se réfère certes à la loi rationnelle, mais applique une jurisprudence plus laxiste, qui tient compte du contexte. Elle assure ainsi le dialogue entre les deux autres niveaux de pensée — exactement comme l'organe judiciaire, dans une démocratie, a pour rôle d'interpréter la loi et de la faire respecter à tous les autres niveaux de pouvoir.

Ces trois chambres que j'envisage — non sans un brin d'humour — ne sont pas sans rappeler la Sensibilité, l'Entendement et la Raison de Kant. J'aimerais les décrire plus précisément, afin de montrer en quoi elles peuvent être rapprochées de la théorie kantienne et en quoi elles s'en séparent.

8.2.1 La Rationalité : les concepts abstraits

La Rationalité, tout d'abord, doit être entendue comme la faculté de raisonner de façon hypothético-déductive. Les jugements rationnels consistent à manipuler des symboles d'après des règles syntaxiques précises, similaires à celles dont j'ai parlé au moment de décrire les théorèmes de Gödel. Ces jugements rationnels sont parfaitement incarnés dans la méthode de programmation top-down, et tout particulièrement dans le modèle des systèmes experts.

Au sein de la Rationalité, il n'y a pas de rapport immédiat avec le réel. Les idées, les données sensorielles, les vécus, les objets, les valeurs sont représentés à l'aide de nombres. La différence essentielle entre cette Rationalité et la Raison de Platon, c'est que la Raison, chez Platon, incluait deux dimensions : la *dianoia* et la *noésis*. Or la rationalité que j'introduis fait l'économie de la *noésis*. Il n'existe pas, dans la conception contemporaine, d'intuition intellectuelle ou de contemplation des idées. Il y a certes, chez Frege, une intuition derrière les axiomes, mais celle-ci se réduit à un sentiment d'auto-évidence — dans tout ce qu'il a de plus cartésien. On peut dire que l'essor de la rationalité moderne se caractérise par une syntaxe détachée de ses contenus sémantiques, où la vérité n'est rien qu'une convention de signes. Les concepts abstraits, eux-mêmes, peuvent être représentés à l'aide de symboles mathématiques. En informatique, le triangle est défini en posant des variables numériques exprimant le nombre de côtés ou la somme des angles.

La rationalité a pour objet — roulements de tambours — les concepts rationnels. Ceux-ci sont définis au moyen de propriétés univoques, le plus souvent vérifiables et quantifiables. Les nombres entiers naturels, par exemple, peuvent être construits au moyen d'un certain nombre de propriétés fondamentales, comme cela se fait dans l'axiomatique de Peano (on en a parlé dans le chapitre 5.2) :

1. L'élément appelé zéro et noté 0 est un entier naturel.
2. Tout entier naturel n a un unique successeur, noté $s(n)$ ou S_n qui est un entier naturel.
3. Aucun entier naturel n'a 0 pour successeur.
4. Deux entiers naturels ayant le même successeur sont égaux.
5. Si un ensemble d'entiers naturels contient 0 et contient le successeur de chacun de ses éléments, alors cet ensemble est $\mathbb{N}$.

Une fois que cet ensemble d'axiomes a été établi, l'intelligence artificielle consiste à développer un système expert susceptible de déduire, de façon automatique, les conséquences qui en découlent. Cette approche top-down permet, dans le meilleur des cas, de découvrir des propriétés arithmétiques qui n'avaient pas été envisagées par l'esprit humain. Elle peut également être utilisée pour compléter (ou simplifier) des preuves concernant des résultats déjà connus.

Un exemple, conçu par Donald Knuth dans les soixante, consiste en un algorithme visant à calculer les n premiers nombres premiers¹⁰ (n pouvant prendre n'importe quelle valeur finie). Pour être valide, cet algorithme ne doit pas se contenter de fournir des nombres premiers, dont l'utilisateur vérifiera la validité *a posteriori*. Dans ce cas de figure, en effet, on quitterait le royaume des concepts purement rationnels, pour entrer dans le domaine de l'expérience¹¹. Ce que l'algorithme doit faire, c'est appliquer une méthode qui soit prouvée rationnellement ; c'est-à-dire, une preuve qui puisse se réduire aux axiomes de Peano. En ce qui concerne l'algorithme de Knuth

10. DONALD E. KNUTH, *The Art of Computer Programming : Fundamental Algorithms*, Addison-Wesley, 1997.

11. Il est en effet recevable de penser que les nombres mathématiques, bien que ce soient des idéautés, peuvent faire l'objet d'une expérience sensible. Un bon exemple est donné par la suite de Syracuse : prenez un nombre entier quelconque ; s'il est pair, divisez-le par deux ; s'il est impair, multipliez-le par trois et ajoutez une unité. En répétant aussi souvent que nécessaire, on observe qu'on finit toujours par revenir sur le nombre 1. Si on commence avec le nombre 7, par exemple, on trouve 7, 22, 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, 4, 2, 1... Cette suite de Syracuse n'est actuellement pas démontrée, et il reste tout à fait possible que, pour un certain entier donné, la suite n'arrive jamais au nombre 1. Jusqu'à présent, cependant, les meilleurs ordinateurs ne sont pas parvenus à trouver un contre-exemple. On est donc face à une expérience numérique qui est valide du point de vue empirique, mais pas du point de vue rationnel. Cela suggère que les preuves formelles sont d'un autre ordre que les vérifications expérimentales.

sur les nombres premiers, une preuve peut être établie, soit au moyen d'un résultat publié au début des années 1850 par Tchebychev, soit en s'appuyant sur une preuve démontrée vers 1934 par Paul Erdős¹².

Les approches top-down, en dépit de leur force de calcul, ne nous enseignent cependant rien de plus que ce qui était déjà connu au moyen des axiomes. Par ailleurs, il peut arriver que ces axiomes soient insuffisants pour décrire parfaitement les objets auxquels ils se rapportent. Les théorèmes d'incomplétude de Gödel, comme on l'a vu dans un chapitre antérieur, montrent que les axiomes de Peano ne permettent pas de démontrer la cohérence de l'arithmétique. Ils sont, dans un certain sens, insuffisants, et ce n'est pas une mince affaire que de contourner cette épineuse difficulté.

8.2.2 La Sensation : les percepts statistiques

La deuxième faculté que j'introduis, ou que je conserve plutôt, est la Sensation. Elle doit être comprise comme la faculté d'établir des associations automatiques, et généralement inconscientes, entre différents vécus sensibles : des impressions, des formes, des sons, des couleurs, des émotions, etc. Il est toutefois important de noter qu'il n'y a pas, au sein de cette sensation, de perception déterminée. On ne peut pas dire de la sensation d'un triangle qu'il s'agit d'un triangle, sans quoi on quitte le registre de la sensation. Le mot « triangle » renvoie en effet à des choses qui ne sont pas présentement dans la sensation : il renvoie à l'idée de toutes les autres choses triangulaires, de tous les triangles possibles et imaginables. C'est pourquoi la sensation ne peut pas faire l'objet d'une définition linguistique. Il n'est pas davantage de prendre la sensation pour objet et de lui donner des prédicats : c'est chaud, c'est coupant, c'est sombre. Tous ces prédicats, une fois encore, renvoient à autre chose que ce qui est senti dans le moment présent.

La Sensation n'existe essentiellement que dans des rapports de similarité et de différenciation. Une forme sensible est un contraste, qui se détache du tout de la sensation, pour acquérir une existence relativement autonome. On ne peut se rendre conscient d'une impression sensible que dans la mesure où on est également conscient de ce qu'elle n'est pas. J'ai illustré cela à plusieurs reprises au moyen des illusions perceptives. Au moment de parler de la chronostase, j'ai argué du fait qu'il n'existait pas d'instant présent, donné d'un seul bloc, de façon objective. L'instant, tout comme l'objet sensible, n'est qu'un découpage réalisé après coup dans un ensemble d'événements temporels voisins amalgamés. C'est pourquoi les jugements sensibles, eux aussi, ne peuvent être que des associations *a posteriori*.

Les jugements sensibles occupent une position analogue aux jugements synthétiques *a posteriori* de Kant. La différence vient de ce que, chez Kant,

12. Ces preuves en questions sont une démonstration du théorème de Bertrand, comme quoi il existe toujours un nombre premier entre deux nombres k et $2k$ donnés (où k est un entier positif quelconque).

le jugement synthétique a posteriori permet de mettre des concepts en relation l'un avec l'autre, comme lorsque l'on dit « cette table est en bois ». Il s'agit certes de *cette* table particulière, et non de toutes les tables, mais il y a néanmoins usage de relations linguistiques. Dans mon concept de jugement sensible, en revanche, les associations sont subconscientes. On pense en effet que cette table est en bois, mais on serait incapable de le justifier ou de l'expliquer. Les objets que l'on manipule ne sont pas des perceptions achevées, mais de simples percepts¹³.

Je ferai l'hypothèse que les percepts, au sein de l'esprit humain, sont définis de façon extensionnelle. Ce que j'entends par là, c'est que le percept d'un arbre, par exemple, n'est rien d'autre que l'ensemble (inconscient) des arbres auquel l'esprit humain a été confronté ou qu'il a imaginé. Il n'existe pas, dans cette catégorie du percept, de processus d'abstraction consistant à extraire, de tous les arbres connus et observables, des propriétés essentielles, comme le fait d'avoir un tronc, d'être fait d'écorces, de racines, de branches, de feuilles, etc. Pour le dire autrement, le percept d'arbre ne s'accompagne pas d'un schéma général, d'une idée floue et caricaturale, d'un symbole d'arbre censé recouper, en lui-même, tous les objets particuliers qu'il désigne.

C'est peut-être la raison pour laquelle le percept est plus ample, plus mouvant, que le concept rationnel ? Il peut arriver en effet qu'on définisse les arbres comme étant dotés de feuilles, puis que l'on découvre soudain un arbre sans feuilles. Si l'on s'en tient au concept rationnel, cet arbre n'en est pas un ; mais du point de vue du percept, ce tronc sans feuille présente de si fortes ressemblances avec son cousin feuillu, que l'on sera porté à les appareiller. Il nous reviendra alors d'étendre le concept rationnel d'arbre en distinguant, au sein de celui-ci, la famille des arbres décidus (qui perdent leurs feuilles en hiver), des semi-décidus (qui perdent leurs feuilles au moment où d'autres apparaissent déjà), des non décidus (qui ne perdent pas leurs feuilles).

Les réseaux de neurones ne manipulent pas de concepts, mais uniquement des percepts statistiques. Il associe des poids spécifiques aux synapses du réseau de façon à pouvoir discriminer une table d'une chaise, ou le bois de la pierre. Il ne donne cependant pas de définition linguistique à ces concepts. Il se contente de dire : « cette donnée que tu me donnes ressemble le plus à celle-ci. ». Il associe donc les nouvelles informations à des classes de données déjà connues. D'ailleurs, chaque nouvel appariement qu'il réalise est susceptible de changer sa structure. Lorsqu'un lien est tissé entre une donnée A et une donnée B, il apprend quelque chose sur A, mais potentiellement aussi sur B. Il peut donc évoluer au cours du temps.

Un logiciel tel que ChatGPT, quoi qu'absolument brillant dans sa maîtrise du langage, manque cruellement de jugement abstrait ? Sauf à lui

13. Deleuze : « le métier du philosophe c'est de faire des concepts, le métier de l'artiste c'est de faire des percepts. » (lettre I, Ve Idée)

donner des règles supplémentaires, il ne peut pas aisément comprendre les concepts qu'il manipule. Il échoue parfois à construire des histoires cohérentes, à jouer aux échecs de façon pertinente, à identifier ses sources ou à s'autocritiquer, se contentant de répéter les suites de mots les plus probables¹⁴

Un réseau de neurones peut être entraîné à manipuler des concepts, même s'il n'en a pas la représentation. S'il reçoit suffisamment de données, il sera capable d'effectuer des opérations arithmétiques et de déterminer des propriétés empiriques en observant des corrélations entre certaines suites de nombres. À titre d'exemple, des méthodes de *Machines Learning* ont été utilisées pour guider l'intuition humaine face à la conjecture de Birch and Swinnerton-Dyer. Cette conjecture décrit l'ensemble des solutions rationnelles des équations définissant une courbe elliptique. Aucune démonstration n'a encore été découverte à ce problème, bien qu'il s'agisse de l'un des sept problèmes de Millénium — récompensé à hauteur d'un million de dollars par l'Institut Clay. Représenter ce problème à l'aide d'un réseau de neurones, c'est décider, plutôt que d'analyser les concepts, de représenter le problème comme une boîte noire associant les entrées et les sorties. De ces corrélations peuvent naître des conclusions utiles. On est passé d'une approche essentialiste à une approche probabiliste.

Il est important de noter que les associations statistiques réalisées par un réseau de neurones ne sont pas identiques aux percepts de l'être humain. La machine structure les données en découpant le réel de façon numérique. Elle segmente une photo en pixels, un enregistrement audio en unités de temps, elle discrétise l'espace pour associer, à chaque point, des valeurs représentant la température, la luminosité, l'albédo, etc. L'être humain, au contraire, ne possède pas d'unités minimales de perception. Il part du tout de la sensation et le découpe de façon à obtenir des formes primitives, qui n'existent que dans le rapport contrasté que ces formes entretiennent avec le tout de la perception. Le jugement sensible de l'être humain n'est donc pas le pur analogon du jugement synthétique de la machine. Il y a toutefois des points communs qui font dire à certains, par une sorte de métonymie mythologique, que la machine est capable de perception.

8.2.3 La Perception : les concepts intuitifs

Le troisième niveau, pris en étau entre les deux précédents, est celui de la Perception. Cette faculté est celle de réunir la diversité des sensations sous une unité conceptuelle que j'appellerai le concept intuitif. La différence

14. Il est cependant à noter que la version o1 de ChatGPT, sortie en septembre 2024, s'est améliorée sur ces aspects. Elle est maintenant capable de créer un fil narratif cohérent du début à la fin. Il y a hélas une architecture de fond qui ne semble pas flexible. Les récits produits par l'algorithme pourraient porter le titre d'un livre de Jean Ségui : « Toujours la même histoire ».

entre ce concept intuitif et les simples percepts de la sensation vient de ce que le concept intuitif est transcendant, tandis que le percept est immanent. Qu'est-ce que je veux dire par là ? Qu'il y a toujours, dans le concept intuitif d'une « table », quelque chose de plus que ce que l'on peut trouver dans chacune des tables particulières. Le concept de table, tout comme le concept de triangle, ce n'est pas cette table-là, ni ce triangle-là. C'est une forme générale, dont les exemplaires concrets ne sont que des représentants incomplets, ou imparfaits. Le concept intuitif est donc proche de ce que Platon nommait l'Idée du triangle, de la marée, du vent hurlant ou de la boue. C'est une unité qu'on se représente au-delà de la sensation, et à laquelle les objets sensibles participent. La différence avec Platon, c'est que ces concepts intuitifs n'existent pas au-delà des choses qu'ils désignent. Il y a ainsi une tension épistémologique entre l'empirisme et l'idéalisme : d'un côté, le concept de triangle n'existe pas ; il ne fait que désigner tous les triangles possibles dans le monde. Mais d'autre part, il est tout aussi vrai de dire que le concept intuitif n'est pas, en soi, aucun des triangles sensibles. C'est une généralisation idéalisée.

Il est toutefois important de noter, et il s'agit d'une autre différence avec la théorie de Platon (voir les chapitres [1.1](#) et [1.2](#)), que cette généralisation idéalisée du triangle, telle qu'on la trouve dans le concept intuitif, n'est pas non plus une idée rationnelle. L'idée rationnelle, en effet, fonctionne à partir de définitions symboliques : elle consiste à dire d'un triangle qu'il est « une figure plane constituée de trois côtés ». Or, lorsque nous manipulons le triangle dans notre intuition, nous n'en avons pas forcément la définition rationnelle. Les études de Piaget indiquent que l'enfant peut se représenter un objet très rapidement, mais qu'il lui faut plus de temps pour donner une structure rationnelle à ces objets qu'il manipule. C'est ainsi que l'enfant détient le concept d'objet dès le stade sensori-moteur, quelques mois seulement après la naissance. Il est capable, non seulement de faire une première forme de distinction entre lui et le monde qui l'entoure, mais il comprend également que les objets ont une certaine permanence. Si son hochet est dissimulé derrière un obstacle, il va contourner cet obstacle pour le retrouver. L'objet ne s'évanouit pas avec la perception qu'il en a. Il existe au-delà du contexte sensible où il est apparu la première fois. Ce n'est donc plus qu'un simple contraste entre les parties et le tout ; c'est une réalité unifiée par une représentation.

J'estime donc que le simple fait de parler d'une « table » ou du « bois » relève déjà d'un autre ordre de jugement que ceux que nous avons vus précédemment : le jugement conceptuel. La Perception est une tentative de réhabiliter l'Entendement de Kant, ce grand « oubli » de la philosophie contemporaine, mais sans faire usage des formes *a priori* de l'intuition et des catégories. Est-ce à dire que je renonce totalement à l'idée qu'il existe des structures innées dans la pensée ? Loin s'en faut. Je cherche néanmoins à faire deux distinctions.

La première, et sans doute la plus facile à se représenter, tient en ceci que les formes *a priori* que je vais défendre ne sont pas des *a priori* au sens le plus universel. Elles ne sont pas nées avec l'homme depuis la nuit des temps, par la volonté du divin créateur. Elles se sont au contraire formées à travers l'évolution des espèces, présentant parfois des écarts importants d'une culture à l'autre. Bien que je sois critique vis-à-vis des prétentions de l'intelligence artificielle à capturer les propriétés de l'esprit humain, je ne nie pas que ces propriétés soient sculptées par l'expérience.

La deuxième différence entre mon concept d'intuition et celui de Kant est plus essentielle : elle vient de ce que l'intuition que je vais défendre ne repose pas sur la théorie euclidienne de l'espace ou la ligne du temps newtonienne. J'ai tenté de montrer, en faisant l'analyse de l'intuition kantienne, qu'il y avait des propriétés plus fondamentales encore que le temps et l'espace. Le temps et l'espace, en effet, se présentent comme le tout de la sensibilité, le réceptacle au sein duquel il ne saurait pas y avoir d'objets ni d'évènements. Il est toutefois douteux que cette forme soit un *a priori* au sens le plus strict du terme. Notre intuition a été sculptée par l'expérience du monde. Notre représentation de l'espace s'est façonnée à la mode de Poincaré et de Lobatchevsky, au travers d'une sommation de sensations et d'expériences sensibles.

Ce qui est plus fondamental que l'espace et le temps, et se tient pour ainsi dire en amont, c'est le fait que les unités minimales de perception, telles que les points de l'espace et les instants du temps, n'existent qu'au travers du Tout de cette perception. Les points n'existent que dans un voisinage de points liés les uns aux autres et les instants n'existent que dans un voisinage d'évènements temporels solidaires. J'aimerais donc substituer à ces intuitions pures de l'espace et du temps, une intuition holistique commune à tous les phénomènes spatiaux et temporels. Cette intuition exhibe une structure. Un objet ne peut pas être pensé en dehors d'un espace qui l'englobe. Cet espace lui-même est formé des autres objets. Il y a donc interdépendance entre les points et l'espace, le tout et les parties.

À cette première forme d'intuition, purement holistique, j'aimerais encore adjoindre une seconde intuition, qui est quant à elle transcendante. L'usage du mot *holistique* renvoie à la capacité à appréhender le tout ; l'usage du mot *transcendant* renvoie à la capacité à concevoir ce qui se tient au-delà du tout tel qu'on peut se le représenter. Qu'est-ce à dire ? J'ai dit que lorsque l'on pensait à un « triangle », on concevait plus que ce triangle-ci ou ce triangle-là. Le triangle, dans son concept, n'est pas un triangle obtus, un triangle aigu ou un triangle rectangle : c'est ce qui est commun à l'ensemble de tous les triangles possibles et imaginables.

Il y a somme toute quelque chose d'extraordinaire, de quasi magique, à cette faculté que nous avons de détacher les idées des objets qu'elles désignent dans le monde à proprement parler, comme si ces idées existaient préalablement à ce monde. Descartes avait bien raison d'insister sur le fait

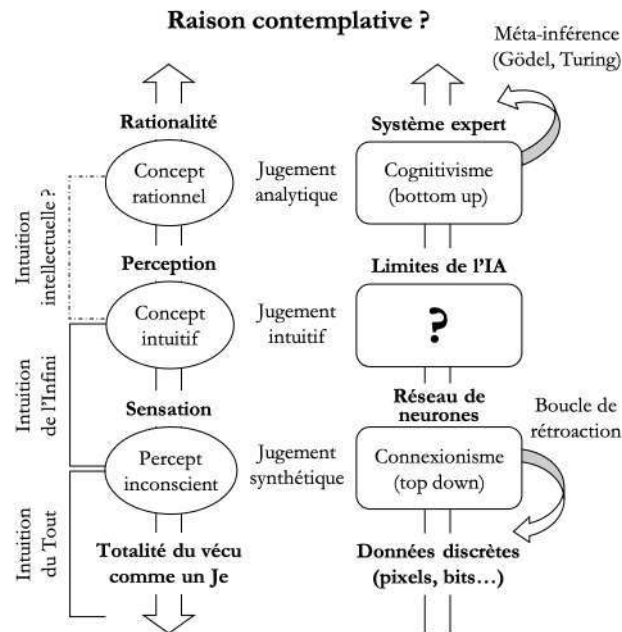


FIGURE 8.4

que des idées sont présentes, en nous, qui dépassent notre capacité à nous les représenter. Il donnait pour exemple l'idée de l'infini. Cette idée n'est visiblement pas quelque chose que nous aurions pu inventer par nous-mêmes. Notre esprit est tout juste capable de mémoriser une dizaine de chiffres (de 0 à 9), que l'on combine assez lentement. Seul Chuck Norris, si l'on en croit la plaisanterie, aurait compté jusqu'à l'infini... deux fois. Mais si quelqu'un osait m'affirmer une telle prouesse, je lui réserverais, en guise de réplique, cette phrase de James Joyce : « Êtes-vous un dieu, un baudet ou un âne bêté ? » Il n'en reste pas moins que lorsque nous contemplons l'horizon, que la mer se déploie sous nos yeux, que les étoiles nous font face, nous ne voyons pas de limite à l'étendue de l'espace et du temps. Quelque frontière que l'on veuille considérer, quelque borne infranchissable que l'on place sur la carte de l'univers, il suffit d'un seul bond de l'imagination pour que le monde s'agrandisse.

Je vais préciser plus exactement la nature et les limitations de ces intuitions holistiques et transcendantes. Mais auparavant, j'aimerais esquisser un schéma général. La figure 8.4 récapitule la théorie tricamérale tel que je défend. On observe que l'esprit humain, représenté à gauche de ce schéma, est doté de trois facultés. Celles-ci partent du « Je pense » cartésien, cet état de perception originel où l'expérience d'être quelque chose plutôt que rien est donnée comme une totalité. Dans cet état, il n'y a pas de différence entre le sujet et l'objet, entre la pensée, la sensation, le vécu et le monde

extérieur. Seul est ce qui est senti, pensé, vécu à l'instant présent. L'esprit humain va toutefois décomposer cette unité pour y faire apparaître des objets indépendants. Ces objets évoluent au cours du temps, ils changent, ils se transforment.

La décomposition de la totalité du vécu en objets individuels n'est pas réalisée d'un seul coup ; c'est un processus qui s'opère par des va-et-vient successifs entre les trois facultés que sont la Sensation (percepts inconscients), la Perception (concepts intuitifs) et la Rationalité (concepts rationnels). Ces niveaux s'appuient l'un sur l'autre : les objets existent au travers du tout et le tout existe au travers des objets. Le fait de passer d'un niveau inférieur de cognition à un niveau supérieur, en partant par exemple du percept pour en faire un concept, est précisément ce que j'ai appelé le *jugement*.

Il existe donc trois types de jugements. Ceux qui consistent à passer de la totalité du vécu (le Je pense) aux percepts inconscients (la Sensibilité) sont les jugements synthétiques et reposent sur l'intuition du Tout. Les jugements qui permettent de rassembler les percepts (la Sensibilité) pour en faire des concepts intuitifs (la Perception) sont les jugements intuitifs et reposent sur l'intuition de l'Infini. Enfin, les jugements permettant d'abstraire les propriétés essentielles des concepts intuitifs (la Perception) pour les manipuler de façon logico-déductive (la Raison) sont dits analytiques et reposent sur une hypothétique intuition intellectuelle (attestée chez Platon mais contestée par Kant, pour qui la Raison spéculative a perdu tout contact avec le monde sensible de l'expérience).

La machine, représentée à droite du schéma, fonctionne différemment de l'esprit humain. Elle ne part pas d'un Je pense originel, mais d'un ensemble préalable d'unités discrètes indépendantes (pixels, points, nombres). Les algorithmes ont pour tâche d'associer ces unités discrètes pour reconstituer une vision globale dans un mouvement purement ascendant (qui part des parties et s'élève vers le tout). La machine ne possède donc que deux niveaux de traitement : (1) celui des réseaux de neurones, qui correspondent chez l'être humain au niveau de la Sensation (paradigme de programmation bottom-up) et (2) le niveau des systèmes experts correspondant à la Rationalité (paradigme de programmation top-down).

Les analyses que j'ai produites tout au long de ce livre tentent de mettre en lumière deux lacunes possibles dans les paradigmes de programmation contemporains. Ces lacunes peuvent être de deux types : (i) l'absence du niveau de la Perception (correspondant à l'Entendement chez Kant) et (ii) l'absence de métacognition (correspondant à la Raison contemplative de Platon), ce qui explique que les algorithmes soient limités par les théorèmes de Gödel et le problème de l'arrêt (Turing).

J'aimerais, avant de conclure cette partie de la thèse, caractériser les deux intuitions que sont l'intuition du tout et l'intuition de l'infini. Il est en effet important d'isoler les ingrédients qui les différencient du fonctionnement de la machine.

8.3 Critique de la séquentialité : L'intuition du Tout

J'ai esquissé les trois dimensions de l'esprit humain que sont la Sensation, la Perception et la Rationalité, mais j'ai gardé sous silence la façon dont il fallait réviser les intuitions kantienne de l'espace et du temps. Au lieu de m'attaquer frontalement à ce problème, je voudrais plutôt insister sur deux ingrédients, propres aux algorithmes contemporains, qu'on ne retrouve justement pas dans l'intuition humaine. Ces ingrédients sont la séquentialité et l'univocité. J'espère qu'en les faisant passer sur le billard épistémologique, le lecteur comprendra mieux pourquoi les machines ne sont pas, ou pas encore, en mesure de devenir de véritables Frankensteins.

Les ordinateurs contemporains fonctionnent de façon séquentielle, en définissant une série d'opérations à exécuter. Ce n'est que lorsque la première est réalisée que la machine peut passer à la seconde, et ainsi récursivement. Même les techniques de programmation dites parallèles, basées sur l'utilisation de thread multiples — pour s'exprimer dans le jargon des programmeurs — ne dérogent pas à cette loi. Elles ne font que juxtaposer différentes séries d'opérations. À ce titre, la machine peut être comparée à une longue chaîne, faite de maillons indissociables, qui se tiennent les uns les autres. À moins d'être lu par une machine de Turing non-déterministe, qui relèverait concrètement de la physique quantique, il n'est pas possible d'effectuer plus d'une seule étape par unité de temps.

Or, lorsque l'on s'intéresse au fonctionnement de l'esprit humain, il n'est pas du tout évident que la pensée s'enchaîne de façon séquentielle. Il semblerait même, au contraire, que la conscience que nous avons d'un certain instant t dépende, dans une certaine mesure, de la conscience de l'instant suivant. Ce n'est qu'une fois que l'instant en question s'est écoulé que nous pouvons, rétrospectivement, le prendre pour objet de pensée, si bien que le processus de pensée se déploie dans un va-et-vient continu entre « ce que je suis en train de penser maintenant » et « ce que je pense que j'étais en train de penser il y a un instant ».

C'est peut-être la raison pour laquelle notre perception du temps est truffée d'interventions ? Lorsque nous écoutons quelqu'un marcher dans le couloir, que ses souliers frappent le carrelage en rythme régulier, nous nous disons : « la personne a d'abord marché et ensuite le bruit est parvenu à mes oreilles ». Or dans notre flux de conscience, nous avons d'abord entendu le son des chaussures, après quoi seulement nous en avons déduit que la personne marchait. Notre raisonnement est donc « il y a du bruit, donc quelqu'un marche ». Ce dernier raisonnement n'est pas *stricto sensu* une déduction logique. Il s'agit bien plutôt d'une abduction, ou encore, d'une recherche de la meilleure hypothèse expliquant l'existence du bruit. Dans notre exemple, on suppose que c'est la personne qui marche qui en est à

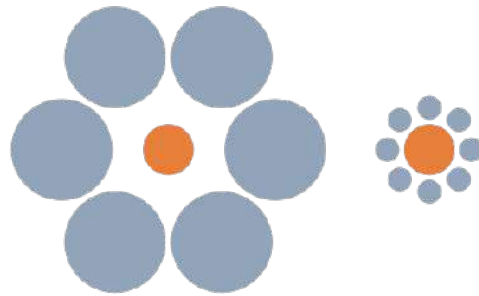


FIGURE 8.5 – L’illusion d’Ebbinghaus illustre qu’un cercle entouré de cercles plus volumineux aura lui-même l’air plus grand.

l’origine, mais ce n’est pas une nécessité logique. En fait, cette recherche de la cause suggère que nous ne pouvons prendre conscience de notre processus de pensée sans inscrire les différents éléments de ce processus (le bruit, une personne qui marche...) dans des relations d’associations logiques ou causales. L’intelligibilité d’un instant particulier dépend de l’intelligibilité du tout auquel cet instant est connecté.

Ce phénomène que je viens de décrire à propos du temps trouve également une contrepartie dans l’espace. La perception d’une scène, peuplée d’objets géométriques, de couleurs, de textures, induit à la fois un balayement des yeux d’un objet vers un autre, mais aussi une intelligence du tout. Ce que j’entends par là, c’est que la disposition globale de la scène, c’est-à-dire les relations que les objets entretiennent entre eux, influence notre perception des éléments individuels. Ce fait est particulièrement bien illustré dans l’illusion perceptive d’Ebbinghaus, dans laquelle deux cercles de même taille sont représentés dans deux situations différentes. Dans la première d’entre elles, le cercle est entouré de cercles plus grands que lui ; dans la seconde, il est entouré de cercles plus petits. On constate que cela influe sur la perception de la taille du cercle originel. Autrement dit, la perception d’un objet isolé dépend de la présence d’autres objets adjacents.

De l’ensemble de ces arguments, il découle que la conscience, loin de fonctionner de façon purement linéaire, implique des jeux d’association qui demandent de réaliser des vas et vient entre l’instant présent, le passé (mémoire), le futur (prédiction), ou entre la cause et la conséquence. Plus rigoureusement, on dira que les notions de passé, de présent, de futur, de cause, de conséquence, sont le résultat de la mise à plat du flux de pensée ; elles apparaissent lorsque l’on tente de structurer les idées traversant notre esprit. Cela ne les rend pas faux pour autant. Ce sont, pour ainsi dire, des conditions de possibilité *a priori*, une ossature sans laquelle il nous serait impossible d’objectifier notre processus de pensée.

8.4 Critique de l'univocité : L'intuition de l'Infini

Une autre propriété des ordinateurs qui mérite d'être mise sur le fil du rasoir, c'est l'univocité. Par « univocité », j'entends un langage qui ne laisse volontairement aucune place à l'ambiguïté. La programmation est fondée sur des instructions logiques, claires et indélébiles. Lorsqu'une instruction est associée à une variable, en écrivant par exemple « $a = b$ », cela signifie que a prend la valeur b . Cette relation est non commutative. Dans un langage de programmation tel que Java ou C++, « $a = b$ » n'a pas la même signification que « $b = a$ ». Dans le premier cas, a est une variable et b est une valeur ; dans le second cas, b est une variable et a est une valeur. Pour donner un exemple concret, une variable peut être, par exemple, la température de la pièce, ou le nom de votre médecin, ou l'argent sur votre compte en banque. Une valeur, quant à elle, est un nombre de degrés Celsius, un nom ou un certain nombre d'euros. On peut donc dire que la température « $T = 26^\circ$ », mais on n'écrira pas « $26^\circ = T$ ». Autrement dit, une température peut être égale à un certain nombre de degrés, mais les degrés ne sont pas en soi une température.

Cette propriété que je décris vient de la définition même des machines de Turing. Cette dernière, en effet, est composée d'un ruban infini, ou disons plutôt indéfini, — prolongeable à l'envi. Sur chaque emplacement de ce ruban peut être inscrit, au plus, un unique symbole. Ce symbole doit être choisi un sein d'un ensemble Σ . L'exemple typique, c'est de ne prendre pour symboles que de 0 et des 1, mais à vrai dire, le choix de ces symboles ne change rien aux propriétés résultantes de la machine, ni à son pouvoir expressif. Elle pourra faire exactement les mêmes choses avec deux symboles ou avec un million, quoi qu'elle sera plus rapide dans ce second cas¹⁵. Toujours est-il que cet ensemble Σ désigne l'ensemble des variables. Dans notre exemple avec la température, il comprend l'ensemble des degrés Celsius possibles, allant de -273,16 degrés à l'infini positif. Dans l'exemple du compte en banque, il va de l'infini négatif à l'infini positif. Et dans le cas du nom du médecin, il comprend les lettres de l'alphabet (plusieurs cases du ruban serviront alors à écrire le nom complet).

On comprend qu'il existe une hiérarchie *a priori*, dans le langage informatique, entre les valeurs Σ et les cellules du ruban (les variables). Cela signifie que, dans le langage informatique, il y a une différence qui existe préalablement entre le sujet (la variable) et le prédicat (la valeur). On dira

15. Cette propriété vient du fait qu'il est possible d'encoder n'importe quel autre symbole de façon binaire. Pour représenter un nombre entier positif, par exemple, il suffit de passer en base binaire. Le nombre 9 s'écrit donc 1001. Il nécessitera quatre cases sur le ruban. Cet accroissement de l'espace nécessaire explique pourquoi la machine se montrera plus lente dans son exécution : il lui faudra parcourir plus de distance sur le ruban avant d'atteindre la fin de son processus.

que la relation de sujet-prédicat n'est pas réversible. On pose d'emblée ce dont on parle comme un sujet sur lequel on vient épingler des attributs bien définis.

Or, il n'est pas du tout évident qu'une telle hiérarchie existe dans l'esprit humain. Lorsque l'on dit d'un individu qu'il est altier ou pingre, on en apprend parfois tout autant sur ce que c'est que d'être altier ou pingre que sur cet individu. Il est possible d'éclairer la nature du prédicat « pingre » ou « altier » en regardant les sujets auxquels ces mots se rapportent. Tout se passe comme si, d'un côté, les attributs possédaient une signification indépendante des personnes auxquelles ils se rapportent ; mais que, de l'autre côté, seule l'existence de gens pingres et altiers donnait une intelligibilité intrinsèque à ces mots. Pour le dire autrement, il y a une relation d'interdépendance entre le sujet et le prédicat. Il est possible d'inverser la relation « Crésus est un pingre » en « Les pingres sont comme Crésus ».

Cette commutativité entre le sujet et le prédicat, spécifique au langage humain, n'est pas sans présenter quelques difficultés. Il signifie en effet qu'il n'existe pas de variables élémentaires et fondamentales, d'unités minimum de significations, sur lesquelles fonder l'ensemble de l'édifice. J'aimerais préciser cela. Dans la machine de Turing, les symboles Σ forment un ensemble irréductible. On suppose que ces caractères ont une signification intrinsèque et autonome, qui ne dépend pas de leur usage dans le programme. Même si un symbole a n'est jamais utilisé, cela ne veut pas dire que a est indéfini. Dans la langue humaine, en revanche, les choses sont bien différentes. Si vous n'avez jamais entendu parler d'un mot, et s'il ne figure pas dans un dictionnaire, vous ne pouvez pas savoir quel est son sens. On ne conçoit les mots français qu'en se référant aux contextes dans lesquels ils apparaissent. C'est pour cela que les dictionnaires sont truffés de référence à des textes classiques. On traque méthodiquement l'usage de tel ou tel terme de la langue française dans l'espoir d'en éclairer les différents usages, les nuances, les évolutions également.

Si ce que je viens de dire est exact, cela signifie qu'on ne peut pas supposer que les mots de la langue française (ou de toute autre langue d'ailleurs) soient définis à partir d'un petit ensemble de mots-axiomes, à partir desquels fonder tous les autres. Elle doit supposer que certains mots sont connus par avance, que leur définition ne dépend pas de la façon dont ils sont utilisés pour définir les autres mots. Or, ce n'est pas forcément la façon dont procède l'esprit humain. Pour nous autres, la sémantique semble être hautement auto-référentielle. Nous ne comprenons les termes du langage qu'en tissant des liens avec la définition des autres mots, qui eux-mêmes renvoient à d'autres mots, qui eux-mêmes risquent de dépendre des mots de départ.

On répondra peut-être que la machine de Turing, si elle devait utiliser les noms du langage, n'a pas besoin de mots-axiomes ; qu'elle peut se contenter d'utiliser des lettres. L'ensemble Σ est ainsi composé des 26 lettres de l'alphabet romain. Lorsqu'elle fait ainsi, ce n'est pas les mots dans leur sé-

mantique et leur définition qu'elle manipule, mais uniquement leur forme syntaxique. Lorsqu'il s'agira d'écrire les définitions de ces mots, elle devra bien supposer que certains mots sont connus par avance, et qu'ils puissent servir dans la définition des autres mots. L'usage des lettres de l'alphabet n'a donc fait que contourner le problème.

C'est d'ailleurs la raison pour laquelle des philosophes tels que John Searle estiment qu'une machine ne comprend pas la sémantique qu'elle manipule. Elle ne fait jamais qu'utiliser des lettres sans signification. À ses yeux, il est égal de manipuler l'alphabet romain ou l'alphabet chinois. Dans un cas comme dans l'autre, elle ne fait que manipuler les règles de la grammaire pour construire des mots qu'elle n'entend guère. Son argument est popularisé par une expérience de pensée connue sous le nom de « l'expérience de la chambre chinoise. » Il imagine un être humain enfermé dans une pièce close, qui aurait pour tâche de communiquer par message avec un interlocuteur à l'extérieur de la pièce. Il ne connaît pas le message, mais il dispose d'une table, certes volumineuse, qui décrit exactement quels symboles il doit écrire face à chaque message qu'il est susceptible de recevoir. Par souci de réalisme, je plaide pour que cet homme soit appelé ChatGPT. John Searle dit à propos de ce chatbot :

Le point de l'argument est le suivant : si l'homme présent dans la pièce ne comprend pas le chinois en implémentant le programme approprié pour comprendre le chinois, alors aucun autre ordinateur numérique ne le fait non plus uniquement sur cette base, car aucun ordinateur, en tant qu'ordinateur, n'a quoi que ce soit de comparable. l'homme n'a pas.

Pour récapituler ce que j'ai dit jusqu'ici, il semblerait que l'être humain, lorsqu'il raisonne, comprenne la signification intuitive des mots qu'il manipule alors même qu'il ne sait pas les définir sans produire des circularités logiques. La machine, de son côté, ne supporte pas la circularité, mais elle ne semble pas forcément comprendre la sémantique des termes manipulés.

J'aimerais illustrer cette haute auto-référentialité inhérente au langage humain en recourant à un exemple concret appelé « le paradoxe du dictionnaire ». Supposons qu'un enfant, vierge de toute préconception, entreprenne de chercher la signification d'un mot compliqué dans un dictionnaire plus volumineux que les almanachs. Admettons que le livre soit suffisamment grand pour embrasser tout le vocable de la philosophie et des neurosciences, de telle manière encore que ces définitions fassent l'objet d'un parfait consensus. Considérons en outre que notre enfant curieux ne connaisse que très peu de vocabulaire, mais qu'il ait la capacité innée d'apprendre des mots nouveaux par le truchement d'une mémoire exceptionnelle. Il connaîtrait la grammaire et ne serait jamais confronté à des difficultés de syntaxe, bien que les théories sous-jacentes aux mots utilisés lui échapperaient complètement.

Il est probable que, rencontrant la définition de « conscience », notre en-

fant avide de savoir, plus proche d'une machine de traitement de texte que d'un être humain de chair et d'os, bute contre de nombreux termes inconnus. La seule chose qu'il pourra dire de ces prédicats, plus mystérieux les uns que les autres, c'est qu'ils sont, tous ensemble, équivalents à la conscience. Si ayant, par exemple, le malheur de tomber sur un dictionnaire phénoménologique, il lisait de la conscience qu'elle était le « flux qualitatif des états intérieurs », il en apprendrait finalement autant sur la conscience que sur le « flux qualitatif des états intérieurs » : à savoir, pas grand-chose de substantiel.

Pour poursuivre sa recherche plus avant, et tenter de dégager la conscience de cette identité monolithique — à supposer, bien sûr, qu'il ne se soit pas découragé du jargon phénoménologique — il lui faudra comprendre ce que c'est qu'un « flux qualitatif » ou un « état intérieur » en se référant à leurs définitions respectives. Hélas, il risque bien, se faisant, de tomber sur de nouvelles phraséologies. Si le dictionnaire est fini, il surviendra alors, tôt ou tard, une circularité, et l'ensemble de la définition se ramènera à dire de la conscience qu'elle est ce qu'elle est. Notre enfant ne l'aura donc saisie, en fin de compte, que dans son unité première et indicible. Si, par contre, le dictionnaire est infini, la démarche peut ne jamais se compléter.

Ce raisonnement par l'absurde suggère que la sensation humaine ne se réduit pas à un acte d'appréhension univoque, du type $A = B$, mais qu'elle se déploie dans une relation d'identités partielles. Chaque fois en effet qu'une chose B est dite à propos d'un sujet A, il faut encore, pour que la relation ait un sens, que cette chose B puisse avoir une existence intelligible en dehors du sujet auquel il se rapporte. Chaque fois qu'un sujet A est défini par rapport à B, il existe donc également, d'après le principe de différenciation, un lien qui dissocie A de B. Or quel peut bien être ce lien en l'absence de tout concept préalable ?

En l'état, la machine psychique ne sait rien produire d'autre que des associations. Il s'ensuit que dissocier un sujet A de son prédicat B ne consiste en rien autre chose que d'affirmer l'existence d'un lien entre B et un troisième élément C qui n'existe pas entre A et B. Chaque fois que deux choses sont mises en rapport l'une l'autre, elles dissimulent ainsi un troisième terme. Celui-ci, comme par l'effet d'un tour de magie, est mystérieusement escamoté du langage formel. Il est pourtant essentiel à l'intuition humaine. Si essentiel, d'ailleurs, que le troisième terme C doit être, à son tour, le prétexte à un quatrième terme D, et ainsi de suite.

À ce stade, sans doute vient-il le moment de me défendre d'une objection évidente. Le fameux principe de différenciation que j'ai introduit, à qui décidément beaucoup de tours de magie sont imputés, ne dissimulerait-il pas une régression infinie ? Tout se produit en effet, dans ce que je viens de raconter, comme dans l'argument du troisième homme d'Aristote.

Cet argument, rappelons-le, est une objection à l'endroit de Maître Platon. Elle consiste à faire remarquer, non sans malice, que si à chaque chose

du monde réel se rapportait une forme dans le monde des idées, alors ces formes devraient être en nombre infini. Il faudra en effet une première forme pour expliquer ce qu'est un homme donné du monde réel, puis une seconde forme pour expliquer le rapport existant entre cet homme du monde et la première forme en question, et ainsi de suite. Cette seule mention d'un cycle infini, dans la pensée des anciens Grecs, exerçait sans doute la même force de dissuasion que l'hypothèse, chez les physiciens contemporains, de l'existence d'une âme. Agitez-la sous leur nez et vous les verrez aussitôt se dédire avec la confusion charmante de l'amoureux devant une bachelette.

Or non seulement je ne pense pas que l'autoréférentiation de la connaissance humaine soit identique à l'argument du troisième homme, mais en plus de cela, et c'est le point le plus important, je ne vois pas en quoi l'infini représenterait une pétition de principe. Au contraire, je pense qu'on peut faire de l'infini une intuition sensible. Une telle intuition se laisse apercevoir, par exemple, lorsque, faisant face à la mer, du haut d'une corniche ou de la proue d'un navire, l'on se représente une étendue d'eau sans fin ni commencement. Inspirant avec nostalgie les embruns crépusculaires, l'on se dit alors avec Baudelaire¹⁶ :

*La mer est ton miroir ; tu contemples ton âme
Dans le déroulement infini de sa lame,
Et ton esprit n'est pas un gouffre moins amer.*

16. *L'Homme et la Mer*, dans *Les Fleurs du mal*.

Le *jugement intuitif* occupe le milieu entre le jugement rationnel de type top-down et le jugements d'expérience de type bottom-up. Il consiste à chercher des contrastes dans le tout de la perception de façon à identifier des objets tels que : une couleur, un pion, de la pierre, une colonne, etc.

Contrairement aux réseaux de neurones, le jugement intuitif ne possède pas d'unités perceptive minimales, tels que les pixels ou les bits. Il ne se contente pas d'associer ces unités les unes aux autres par le jeu des lois statistiques ; sa tâche est, tout au contraire, de construire des unités d'analyse.

Pour le dire autrement, l'esprit humain est composé de trois niveaux cognitifs : la sensation (bottom-up), la perception (jugement intuitif) et la rationalité (top-down). Le passage allant de la sensation à la perception est rendu possible au moyen de deux ingrédients :

1. *L'intuition de l'Infini* est la capacité à penser le rapport existant entre un concept universel, tel que le concept d'un « pion », et la multiplicité des pions singuliers désignés par ce concept.
2. *L'intuition du Tout*, quant à elle, est la capacité à penser de façon non séquentielle, en saisissant toute l'interdépendance des concepts et des objets.

Ces intuitions sont l'analogue, chez Kant, des intuitions de l'espace et du temps. Ma théorie tricamérale de l'esprit se présente ainsi comme une révision en profondeur du jugement synthétique *a priori*.

Dans la prochaine partie de cette thèse, je vais tenter de caractériser plus précisément ces deux intuitions, et montrer en quoi elles échappent à nos ordinateurs.

Le Jugement intuitif face à la Machine

Extension de la machine de Turing

Chapitre 9

Le Dilemme de l'intuition

*L'intuition est ambiguë, elle
renferme en elle une infinie
complexité, et montre plusieurs
faces les unes après les autres.*

Arthur Schopenhauer

Une machine peut-elle penser ? Il est grand temps de se demander ce que les outils conceptuels que j'ai adoptés dans les précédentes parties peuvent nous enseigner à propos de la question qui fait l'objet de cette thèse. Je me suis appliqué à justifier, non sans diligence, qu'il existait trois couches à la pensée humaine : le niveau des percepts (sensation), le niveau des concepts intuitifs (perception) et le niveau des concepts rationnels (concepts abstraits).

Les concepts abstraits, ou rationnels, sont définis uniquement à l'aide de mots ou de nombres. Ce sont de purs signes syntaxiques. Ils ne capturent pas toujours une réalité sensible et expérimentable, mais peuvent être manipulés par le truchement du raisonnement logique. C'est ce niveau de pensée qui semble le plus proche du fonctionnement de l'ordinateur, mais c'est également la dimension la plus superficielle. Au moment de parler des théorèmes d'incomplétude de Gödel, j'ai indiqué que, dans une théorie axiomatisée un tant soit peu complexe, la majorité des propositions sont syntaxiquement indémonstrables : on ne peut pas dire si elles sont vraies ou fausses sur la seule base des axiomes. Les rares vérités qui succombent à la logique sont des points isolés, des îles solitaires, perdues au milieu d'un océan infini. La raison calculante, pour le dire autrement, est la peau qui recouvre le fruit, l'écorce qui dissimule la sève !

Le niveau des percepts, quant à lui, est de l'ordre du sensible. La perception s'y présente comme un tout, un voisinage d'évènements et de formes spatiales amalgamées. Il n'y a pas d'unités minimales de la perception, telles que des pixels ou des instants, qui servent de points de référence. En l'ab-

sence de ces briques élémentaires, l'esprit doit isoler des contrastes au sein de la totalité perceptive, en suivant un principe consistant à différencier et à assimiler les parties les unes aux autres. Des formes naissent, qui ne sont pas toujours conscientes. Ce sont tout au plus des sensations fugitives, parfois intenses, mais fortement liées au contexte de l'expérience, où il n'est pas permis de distinguer aisément le sujet qui éprouve de ce qu'il éprouve. On ne peut pas isoler le percept et dire qu'il a une réalité en dehors de l'expérience sensible elle-même. Ce niveau du jugement que je décris est celui que tentent d'imiter les réseaux de neurones. Des corrélations statistiques sont établies de sorte à faire naître des catégories au sein des données sensibles. Bien que les résultats soient bons, je me suis efforcé de montrer que la façon dont ces algorithmes représentent les données diffère de la perception humaine.

Le troisième niveau, situé entre les deux premiers, est celui des concepts intuitifs. Ici, l'esprit humain isole une forme perceptive et lui donne une réalité intelligible, consciente, exprimable linguistiquement. Dans le processus s'établit une distinction entre le concept et la chose que ce concept désigne. Il est possible d'adopter un point de vue platonicien, dans quel cas le concept existe de façon idéale et les formes sensibles n'en sont que des expressions particulières. Mais il est également possible de prendre le point de vue empiriste, dans quel cas ce sont les formes sensibles qui sont premières et nos concepts ne font rien d'autre que de les synthétiser. Dans un cas comme dans l'autre, le concept est une réalité manipulable par la pensée, mais liée à un contenu sensible. Il n'est ni une pure abstraction syntaxique (comme dans le niveau de la Rationalité), ni une simple sensation singulière (comme dans la Sensation).

J'ai suggéré que les concepts intuitifs de ma théorie occupaient, vis-à-vis de la pensée kantienne, une position analogue aux jugements synthétiques *a priori* du type : « La somme de 5 et 7 fait 12 » ou « La somme des angles d'un triangle fait 180 degrés ». Ces derniers, en effet, ne se fondent ni sur la raison pure, ni sur l'expérience pure, mais sont des conséquences de la façon dont l'être humain structure les phénomènes. La grande idée de Kant, pour l'exprimer dans toute l'élégance de l'art, c'est que les données sensibles ne sont pas perçues telles qu'elles existent en soi, mais qu'elles prennent des formes qui dépendent de nos concepts et intuitions préalables. Ces derniers appartiennent au bagage inné de l'être humain. Lorsque nous pensons que l'Univers est un espace euclidien ou que le temps est une réalité objective, nous nous comportons à la façon d'un ingénieur zélé qui, après avoir apposé partout la marque de ses outils, pense que la nature a fait le monde avec les mêmes outils que lui.

Si la machine ne pense pas comme l'être humain, c'est parce qu'elle ne structure pas les phénomènes sensibles avec les mêmes outils que nous autres. Pour s'en rendre compte, voyons d'abord comment les choses se présentent dans la perception humaine, puis comparons-les à nos algorithmes. Au sein de notre esprit — je pense l'avoir dit plusieurs fois — il existe un rapport

étroit entre le tout de l'espace et les parties de celui-ci. Les points de l'espace ne sont pas des données brutes ; ils tirent au contraire leur intelligibilité du lien entretenu avec les autres points du voisinage. Il y a ainsi une imbrication entre l'espace, pris dans son ensemble, et les figures qui se déploient en son sein. C'est ce qui faisait dire à Kant que l'espace est le réceptacle des phénomènes, qu'il est ce sans quoi aucun corps ne peut être pensé.

Un phénomène analogue, quoique plus complexe à caractériser, se produit pour la perception humaine du temps. Les événements ne sont pas donnés sur une ligne chronologique allant du passé vers le futur. Le cerveau humain — auquel on impute bien des tours de magie — ne se contente pas de placer les stimuli à la file indienne. Aux yeux de l'esprit humain, il n'y a pas d'instants en soi, mais uniquement les changements intervenant sur des intervalles de temps. C'est pourquoi le temps subjectif dépend, pour nous, de la quantité de changement survenant dans nos vécus intérieurs : lorsque le paysage change, que nos pensées vagabondent, que nous rêvassons, que nous éprouvons des émotions fortes, le temps passe plus lentement que lorsque nous restons dans une absolue fixité.

Or les algorithmes recourent à une toute autre représentation des données sensibles. Ils disposent pour commencer d'une horloge interne ; un certain mécanisme qui leur permet d'effectuer, à chaque unité de temps, exactement une opération numérique. De plus, la machine ne peut pas penser les rapports liant le tout et ses parties de façon interdépendante. Elle découpe l'information en unités de perception minimales : une photo est décomposée en pixels, un enregistrement audio est digitalisé, les mots sont décomposés en lettres et lues l'une après l'autre (là où l'être humain est attentif au mot dans son entier¹). L'ontologie de la machine est un flatland, un monde plat, sans relief, où les savoirs s'obtiennent par juxtaposition de briques élémentaires. Pour le dire en une phrase lapidaire : les structures de perception humaine sont holistiques, tandis que les structures de la machine sont discrètes.

Dans cette troisième partie de ma thèse, j'aimerais reprendre deux questions que j'ai laissées dans l'ombre. La première consiste à savoir comment exactement se définissent l'intuition holistique et l'intuition de l'infini, dont j'ai prétendu qu'elles étaient au fondement de la perception humaine. La seconde consiste à se demander ce qu'il faudrait changer au sein des machines que nous utilisons actuellement pour leur permettre d'avoir une intuition humaine. Ces deux questions se rejoignent sur un objectif commun : formaliser la notion d'intuition pour lui donner une forme mathématiquement vérifiable. Si, en effet, j'arrive à représenter ces intuitions à l'aide d'une fonc-

1. Une étude de l'utrivsineé de Crbindage a mrtnoé que l'on peut snas pbmlèore lrie un texte dnot les lrtets sont dnas le drrésode pour peu que la prierème et la dirnrèee lttere de cauqhe mot rtenest à la bnone pacle. Ceci motnre que le careveu ne lit pas teuots les lteetrs mias pnred le mot comme un tuot. La pveure : avouez que vuos n'avez pas eu de mal à lire ce ttexe. (Pour un commentaire de ce paragraphe, voir THIERRY OPILLARD, *La Dégradation Cambridge : Même pas vrai ?*, Les Actes de Lecture n.102, juin 2008.

tion mathématique, il serait aisé de voir si cette fonction est, oui ou non, calculable par une machine de Turing. Il suffirait en effet de déterminer si, en donnant la bonne entrée sur le ruban, elle est capable de donner la bonne sortie.

Cette première formulation de la question est toutefois trompeuse. Elle suppose en effet que l'intuition soit représentable par une fonction et que celle-ci puisse être exprimée dans le langage de la machine de Turing. Or cette dernière, on l'a dit, se représente l'information en la discrétisant. Il n'est pas sûr qu'elle puisse jamais posséder, dans ces conditions, l'intuition d'un point indivisible. Je vais expliquer plus soigneusement cette objection.

Lorsqu'il dessine au crayon un point sur la feuille, l'être humain sait que ce point grossier contient en lui-même une infinité de points mathématiques. Il le sait, car il décompose spontanément le tout en parties ; il lui serait même impossible de ne pas poursuivre, au moins par imagination, ce processus de décomposition. L'ordinateur, au contraire, part du principe qu'il existe un point, le pixel, qui est indivisible.

Cette quête qui consiste à définir l'intuition nous conduit invariablement face à la question du continu : un ordinateur peut-il penser une forme, un segment, une droite, dans son ensemble ? Peut-il se représenter comme l'être humain le rapport qui unit cette forme, dans son tout, à chacune de ces parties ? Ou ne peut-il se représenter l'espace que comme l'agglomération de pixels indivisibles ?

Ces interrogations préliminaires me conduisent à une première hypothèse, que je tenterai de mettre à l'épreuve tout au long de ces pages :

Hypothèse 4. *Seule une machine de Turing capable de penser le continu ou de penser le tout pourrait avoir accès à l'intuition holistique et à l'intuition de l'infini qui caractérisent la perception humaine.*

9.1 L'Intuition de l'Infini et l'intuition du Tout

La machine de Turing est le modèle par excellence de l'ordinateur, dont j'ai donné une description précise dans la section [5.3.1](#). Elle consiste en un ruban de taille infinie, sur laquelle une tête de lecture inscrit des suites de symboles du type 1 ou 0. Cette tête, qui n'est rien d'autre que le stylet sacré de l'informatique, peut lire les symboles existant, les effacer, en écrire de nouveau et se déplacer autant qu'elle le souhaite le long du ruban. La question que j'adresse consiste à se demander si ce modèle de Turing serait capable de construire des concepts intuitifs ? Ce qui a été exposé dans la partie précédente de la thèse me permet de relever au moins quatre obstacles théoriques :

- **Discrétisation :** Au sein d'une machine de Turing, le temps est composé d'une succession d'instants et le ruban est fait d'une succession de cases. C'est ainsi qu'à chaque unité de temps, il est possible de

réaliser exactement une opération parmi : se déplacer d'une case vers une autre, lire un symbole ou changer un symbole. Or dans l'esprit humain, il semblerait que les points de l'espace n'existent pas préalablement à l'acte de perception. J'ai démontré, au moyen de plusieurs illusions, que l'esprit appréhende le tout, non en le grillageant centimètre par centimètre, mais en identifiant les contrastes existant entre ce tout et certaines de ses parties. Lorsqu'il est mis face à une photo, il ne commence pas par balayer les pixels de gauche à droite et de haut en bas. Son regard suit une trajectoire certes prédictible et orientable, mais qui dépend des contrastes et des saillances au sein de l'image. Les points ne sont pas des réalités individuelles indépendantes ; ils tirent leur intelligibilité du tout auquel ils se rapportent.

- **Séquentialité** : La machine de Turing possède une horloge interne. Elle effectue des opérations de façon séquentielle, en partant d'un point A pour aboutir à un point B. Or j'ai cherché à défendre que le temps, tel qu'il est vécu par l'être humain, n'est pas ordonné de façon absolue. Les expériences menées sur la chronostase montrent que, lorsque deux stimuli sont espacés de moins de 100 millisecondes, il peut arriver que nous les intervertissions. Ce phénomène n'est pas un choix délibéré de la volonté. Il est exécuté de façon automatique lorsque certaines conditions sont rencontrées. Tout nous porte à croire que la perception humaine ne se porte pas successivement sur des instants bien définis, mais sur des voisinages d'événements temporels. Elle est plus sensible aux contrastes existant entre les phénomènes de ce voisinage qu'aux durées absolues qui les séparent.

- **Univocité** : La machine de Turing est définie par des assignations univoques. Lorsqu'on écrit $A = a$, cela signifie que la valeur a est assignée à la variable A , ou autrement dit, que la machine de Turing a inscrit le symbole a sur la case A du ruban. Cette notion semble constitutive de la calculabilité. Un calcul ne consiste justement pas à trouver la valeur d'une certaine variable ?

Or dans l'esprit humain, il n'est pas sûr que la relation de prédication soit univoque. La proposition « ceci est un arbre » nous apprend quelque chose de significatif sur la nature du « ceci » qui se tient face à nous, mais également sur ce que c'est que d'être un « arbre ». Le sujet de la proposition (ceci) et sa propriété (être un arbre) sont commutatifs.

Autrement dit, ce n'est pas en lisant la définition d'un arbre dans le dictionnaire que nous nous faisons une idée correcte de ce dont il s'agit ; c'est en se promenant dans la forêt, en comparant les différents troncs, les feuilles, les ramures, en s'exposant à toute la multiplicité sensible renfermée dans le concept. Cela implique que lorsque

nous pensons au concept « d'arbre », nous pouvons nous représenter une série d'arbres distincts les uns des autres. Ces arbres partagent entre eux des similarités, mais aucun n'est parfaitement identique. On pourrait donc penser que l'être humain, lorsqu'il construit un concept aussi banal que celui d'un « arbre », manipule des possibilités infinies. Il synthétise cet infini sous une image mentale unique.

- **Absence de sémantique :** Cette capacité de l'être humain à appréhender l'infini se fait sentir lorsque nous nous représentons l'ensemble $\mathbb{N}$ de tous les entiers naturels. Nous ne pouvons pas énumérer explicitement les entiers 1, 2, 3, ... Notre mémoire est bien trop limitée pour en faire ainsi. Il existe néanmoins, dans notre esprit, une faculté proprement extraordinaire qui nous permet de généraliser la séquence de nombres pour tendre, asymptotiquement, vers un nombre toujours plus grand. Cela se produit par un procédé d'analogie.

Supposons que nous voulions nous faire une idée du nombre mille, du nombre 150 millions ou du nombre 1 suivi de 17 zéros, de façon à nous approcher tendanciellement de l'infini. Pour ce faire, nous commençons par nous représenter une bille d'un centimètre placée au centre d'un terrain de foot. Nous constatons que le nombre mille équivaut au rapport existant entre le diamètre de cette bille et la taille du terrain de foot. Nous pensons ensuite que si cette bille était le soleil, les 150 millions de kilomètres séparant la terre du soleil seraient comparable à la distance séparant cette bille du bord du terrain de foot. Nous songeons ensuite que les 28 000 années-lumière (soit 10^{17} kilomètres), séparant le soleil du centre de la voie lactée sont comparable à la distance séparant notre bille du soleil. En répétant cette expérience de fil en aiguille, de sorte à nous représenter des distances toujours plus grandes, nous arrivons à l'idée vertigineuse de l'infini.

La faculté de concevoir en imagination une idée non représentable, voilà bien quelque chose qui semble échapper à la machine de Turing ! Cette dernière possède certes un ruban de taille infinie sur lequel inscrire les nombres 1, 2, 3, ..., mais peut-elle se faire une idée de ce qui existe au bout de cette séquence ? Lorsqu'elle écrit le symbole $\mathbb{N}$, qu'elle énumère ses propriétés et le manipule au sein d'une démonstration, elle se contente d'une définition rationnelle, abstraite, purement linguistique. Elle comprend la syntaxe, mais qu'en est-il de la sémantique ? Ne se comporte-t-elle pas comme un singe manipulant des mots sans comprendre ce qu'ils dénotent ? C'est certes un singe précis et exact, un singe savant, mais cet esprit est-il doté de l'Entendement humain ?

Dans les pages qui suivent, je ferai l'hypothèse qu'il existe deux formes

d'intuitions : l'intuition holistique, qui concerne le rapport entre le tout et les parties, et l'intuition transcendante, qui se rapporte à l'infini. Ces dimensions de l'intuition sont l'enclume sur laquelle l'esprit humain, muni du marteau de l'expérience, peut forger les catégories. Sans cette enclume, l'expérience brasserait de l'air. Il lui faut une résistance pour donner naissance à des objets. L'intuition, et ses facettes, est donc la faculté à se saisir des objets, non pas à l'aide d'un raisonnement logico-déductif traçable, qui procède linéairement des prémisses vers les conclusions, mais de façon globale. Si l'on en croit l'étymologie latine, *intuition* viendrait d'ailleurs du mot *intuitio*, qui signifie « l'acte de voir d'un seul coup d'œil. » Il y a d'ailleurs une tendance, en philosophie, à opposer le concept d'intuition à celui de raison. Il suffit de penser par exemple à cette phrase de Schopenhauer : « Les concepts forment une classe spéciale de représentations, entièrement distincte des représentations intuitives dont il a été question jusqu'ici, car elles n'existent que dans l'esprit humain² »

La première forme d'intuition, *l'intuition holistique*, appelée aussi l'intuition immanente, peut être désignée comme une certaine faculté à comprendre les relations d'impermanence existant entre les concepts. Là où l'esprit rationnel pose les mots à plat, en tentant de construire entre eux des relations hiérarchisées, l'intuition immanente, au contraire, fonctionne de façon cyclique. La raison simplifie le problème, fixe les éléments de départ, déduit des conclusions à la façon de Descartes dans son Discours de la Méthode ; l'intuition complexifie, tresse des relations et ouvre des perspectives plus qu'elle n'apporte de réponse. En psychologie, cette intuition immanente est parfois désignée sous le nom de « pensée en arborescence », de « pensée divergente » ou de « cerveau gauche ». Ces termes revêtent cependant des définitions variables, quelque peu flottantes, et à l'exception du second, ils ne sont pas établis dans la littérature scientifique. Quelqu'un disposant de cette intuition de façon exacerbée aura tendance à se défier des concepts, qui lui sembleront insuffisants pour saisir la richesse et la densité des relations entre les choses. La seule idée de décomposer le monde en objets disposant d'une existence individuelle, plutôt que de se le représenter comme un tout, lui semblera douteuse.

Cette intuition holistique est l'ingrédient principal de la Sensation, lorsqu'elle produit des percepts. Ces derniers en effet n'existent pas de façon indépendante. Ce ne sont pas des objets individuels stables, identifiables en dehors de leur contexte, mais des contrastes qui apparaissent au sein d'un tout. C'est d'ailleurs la raison pour laquelle, lorsque l'on vous présente des taches de café sur le papier blanc, similaires au test de Rorschach, des formes apparaissent dans votre esprit (voir figure 9.1). Vous voyez tantôt un visage, tantôt un monstre, tantôt des personnages, des montagnes. Ces contours n'existent cependant que de façon indéterminée. Si vous changez l'arrière-

2. Schopenhauer A., *Le monde comme volonté et représentation*, Chapitre 9

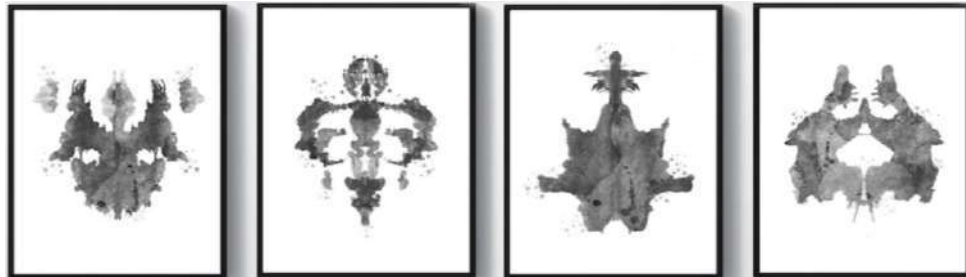


FIGURE 9.1

plan ou que vous tournez la feuille, elles sont susceptibles de se transformer. C'est en raison de cette intuition holistique que la structure de la perception humaine, contrairement à la machine de Turing, ne semble pas s'inscrire dans une représentation discrète et séquentielle du monde.

La seconde forme d'intuition, *l'intuition de l'infini*, appelée aussi l'intuition transcendante, est la faculté à se représenter l'irreprésentable ; à comprendre que les symboles désignent autre chose que ce qui est contenu en eux. À l'inverse de la rationalité, qui manipule les signes du raisonnement, qui agencent les mots d'une phrase comme un gramophone agence les phonèmes, l'intuition transcendante comprend que l'image mentale associée au concept arbre n'est pas le concept proprement dit. C'est ce qui lui permet d'opérer une distance entre les mots et les choses, entre les signifiants et les signifiés. Cette aptitude se manifeste en particulier dans la possibilité de saisir des idées, comme l'idée de l'infini. Il est d'ailleurs curieux qu'une telle idée apparaisse en nous dès le plus jeune âge, que l'on songe tantôt au monde extérieur (les corps se déploient dans un espace infiniment extensible) ou tantôt au monde de notre conscience (les phénomènes se manifestent dans une temporalité infiniment extensible).

L'intuition transcendante, en toute logique, est déjà présente, de façon embryonnaire, au stade des percepts. En effet, ces percepts sont constitués d'un nombre infini potentiel de parties. Les taches de Rorschach peuvent être divisées en sous-taches, il est possible de voir des formes au sein des formes initiales. L'infini ne se manifeste toutefois pleinement que lorsqu'elle parvient à donner à ces catégories une image unifiée en disant de ces formes : « c'est un homme ! » ou « c'est un arbre ! ». C'est alors en effet que surgit l'idée que le concept « d'arbre », en tant que concept intuitif, déborde toutes les perceptions d'arbres connues et imaginables. C'est en raison de l'intuition transcendante que l'esprit humain peut donner un sens, une sémantique, aux symboles qu'il manipule³.

3. Il s'agit d'une remarque purement technique, mais il peut être judicieux de faire remarquer qu'il est possible de contourner les théorèmes d'incomplétude de Gödel en recourant à une preuve transfinie. Tout se passe donc comme si l'être humain pouvait donner du sens à certaines propositions indémontrables parce que sa sémantique, ou son expé-

9.2 L'Objection d'Alan Turing

C'est une chose de dire qu'une machine de Turing n'aurait pas d'intuition, mais c'est autre chose que de le montrer. À partir du moment, en effet, où l'intuition se définit en opposition avec la réflexion logique, il peut être difficile d'explicitier rationnellement en quoi elle consiste. Or si on ne précise pas, en des termes rigoureux, ce qu'est cette intuition, mais qu'on dit de façon vague, par exemple, « La machine ne peut pas avoir du génie », on est reconduit à la question de savoir ce qu'est le génie. Ne serait-ce pas un concept fourre-tout, au même titre que les idéalistes continuent de voir, dans l'âme, l'ultime bastion face au matérialisme ? Dans son célèbre article, *Computing Machinery and Intelligence*, Alan Turing avait déjà anticipé ce problème⁴ :

Ces arguments prennent la forme : « Je vous accorde que vous pouvez faire faire aux machines toutes les choses que vous avez mentionnées, mais vous ne pourrez jamais lui permettre de faire X. » De nombreuses fonctionnalités X sont suggérées à ce propos. Je vous propose une sélection : être gentil, débrouillard, beau, amical, avoir de l'initiative, avoir le sens de l'humour, distinguer le bien du mal, faire des erreurs, tomber amoureux, déguster des fraises et de la crème, faire que quelqu'un en tombe amoureux, apprendre par expérience, utiliser les mots correctement, être le sujet de sa propre pensée, avoir autant de diversité de comportement qu'un être humain, faire quelque chose de vraiment nouveau.

Or parmi ces choses X qui sont énoncées, il y en a beaucoup qui semblent vagues ou difficilement définissables. N'est-ce pas en raison de ce flou conceptuel que l'on est tenté de penser que la nouveauté, par exemple, résiste aux algorithmes ? Si on fournissait un critère mathématique permettant de déterminer si une production artistique donnée (un dessin, une poésie, une musique...) est éloignée statistiquement d'un ensemble de productions connues, alors il n'y aurait pas de raison de penser qu'un réseau de neurones ne puisse pas être entraîné à produire de la nouveauté. Il se pourrait même que la nouveauté de la machine surpasse rapidement l'être humain.

Mais je crois déjà entendre mes détracteurs s'écrier dans leur barbe philosophique : « la nouveauté ne se réduit pas à une pure et simple indépendance statistique, mais repose sur un autre ingrédient primordial ». Quel est cependant cet autre ingrédient ? Le génie ? L'originalité ? Le changement de

rience du monde, n'est pas limitée par une représentation finitiste. Ce point nécessiterait cependant de plus amples développements. J'aurai l'occasion d'en parler plus longuement lorsque je définirai plus exactement les machines de Turing transfinies

4. La traduction est de moi. Le passage vient de la section « Arguments from Various Disabilities ».

cadre ? Dans notre tentative d'expliquer le X de la nouveauté en des termes vérifiables, nous nous sommes heurtés de plein fouet à un X nouveau et mystérieux.

Ce problème est ce que j'appellerai le *dilemme de l'explicitation*. Il s'exprime de la manière suivante : si l'on suppose qu'une certaine faculté (comme l'intuition) marque une distinction entre l'esprit humain et le fonctionnement d'une machine, alors

- Soit cette faculté est définie en termes explicites et vérifiables, dans quel cas rien ne nous interdit de croire qu'une machine est susceptible de l'imiter ;
- Soit cette faculté n'est pas définie en des termes explicites et vérifiables, dans quel cas on peut douter de son existence.

Notons que ce dilemme n'est pas exclusif à l'intuition, mais concerne toute faculté susceptible d'être invoquée pour dessiner une ligne de démarcation irréfragable entre l'être humain et la machine : qu'il s'agisse de la mémoire, de l'émotivité, de la subjectivité, etc. L'intuition a tout de même ceci de particulier qu'elle désigne, dans la plupart des définitions, une aptitude à connaître qui s'oppose aux concepts de la raison. Elle se pose ainsi comme un point de résistance privilégié à l'idée que l'esprit, par nature, ne soit rien d'autre qu'un ensemble d'opérations calculables. Cette opinion est notamment illustrée par Schopenhauer :

Le fonds de la vraie sagesse, la vraie science de chacun n'est point enfermé dans des concepts, dans un savoir abstrait, elle est tout entière dans l'intuition et dans le degré de pénétration, de justesse et de profondeur avec lequel il a saisi cette intuition.

Remarquons également que certaines définitions de l'esprit humain ont le mérite d'être explicites, quoique non vérifiables. C'est le cas lorsqu'il est dit que seul l'être humain est capable de sensations, ou autrement dit, qu'il possède en lui un monde intime. Statuer de l'existence de cette intériorité chez un être vivant ou un être artificiel ne semble pas pouvoir recevoir une réponse évidente ; et les gens exposent sur ce genre de matière des opinions généralement fort différentes. C'est pourquoi je chercherai d'abord à parler de l'intuition en des termes vérifiables pour me demander, à l'issue de cette première partie, si les conclusions obtenues sont en mesure d'enseigner quelque chose quant aux problèmes métaphysiques plus fondamentaux.

Il n'est pas étonnant, au vu de ce dilemme, d'observer un cloisonnement du débat à propos de l'intelligence artificielle : les optimistes arguent du fait que de nombreuses tâches, soi-disant hors de portée de la machine, ont finalement été accomplies par celle-ci — et que les quelques tâches laissées ouvertes ne manqueront pas de suivre le même phénomène. Les pessimistes, quant à eux, développent des définitions qui, par leur nature, ne sont pas

assez mathématiques que pour être mises à l'épreuve de la machine.

Pour dépasser ce dilemme, la philosophie est mise en demeure d'ouvrir une tierce voie. Celle-ci consisterait à donner une définition de l'intuition qui, bien que parfaitement explicite, ne serait toutefois pas de l'ordre du calculable. Cette idée, ainsi exposée, peut paraître un peu étonnante. Loin de consister en une simple abstraction, il est toutefois possible d'en donner une analogie mathématique : en analyse, il existe des équations différentielles sans solution analytique. Il est possible de montrer que ces équations ont une solution exacte, mais il est impossible de la synthétiser sous une formule facile à calculer ; si bien qu'un ordinateur ne peut, tout au plus, qu'en donner des approximations.

Ce genre d'équations est même si fréquent, en science, que les solutions exactes constituent une exception. En physique classique, par exemple, l'effet de la gravitation sur un système composé de plus de deux corps ne peut pas être calculé parfaitement ; et en physique quantique, les équations de Schrödinger ne sont parfaitement solubles que pour le seul atome d'hydrogène. À supposer que la métaphore s'applique à l'esprit humain, et que celui-ci soit plus complexe qu'un atome d'hydrogène, se pourrait-il que certaines des intégrales intervenant dans les processus de l'esprit humain, quoique parfaitement définies, soient néanmoins hors de portée du calcul ?

La solution que je vais élaborer dans cette troisième partie de ma thèse consistera à concevoir un formalisme mathématique permettant d'exprimer l'intuition de l'infini. Ce formalisme devra être le plus opérationnel que possible, mais tout en échappant au fonctionnement d'une machine de Turing classique. Je montrerai qu'il peut également servir pour entrer en dialogue avec les neurosciences, particulièrement en ce qui concerne la construction d'analogies de structure entre le fonctionnement du cerveau et les propriétés de l'expérience phénoménologique. Mais avant de me lancer dans cette tâche ardue, je vais tout d'abord montrer que certaines définitions de l'intuition que l'on trouve dans la littérature, et particulièrement celles qui ont été élaborées en réponse aux problèmes soulevés par l'intelligence artificielle, ne résistent pas au dilemme de l'explicitation.

Dans ce chapitre, je vais discuter la définition de Hubert Dreyfus, qui a eu un impact significatif sur les paradigmes de recherche en *intelligence learning*. Je montrerai en quoi cette définition opère une distinction intéressante entre le savoir (know-that) et le savoir faire (know-how). Je soulignerai cependant que, depuis que les algorithmes se sont saisis des réseaux de neurones, de la programmation évolutionnaire et d'autres méthodes d'apprentissage automatique, il est aisé de voir que les machines sont capables d'implémenter cette définition de l'intuition.

J'envisagerai alors deux stratégies classiques de la littérature pour se sortir du dilemme de l'explicitation. La première, qualifiée de *mystérianisme*, consiste à dire que l'intuition existe, mais qu'elle n'est pas définissable au moyen de notre intelligence humaine. Il s'agit alors tout simplement de nier

la pertinence du dilemme de l'explicitation. Cette position doit être étudiée avec soin ; car s'il était avéré, non seulement toute démarche qui tenterait de définir l'intuition serait vouée à l'échec, mais il se pourrait en outre qu'elle invalide partiellement les résultats que j'ai obtenus au travers de ma théorie méréologique de la conscience. Et pour cause, lorsque cette posture philosophique a été popularisée sous la plume de McGinn, elle ne concernait pas l'intuition, mais la conscience elle-même.

La seconde stratégie que je souhaite aborder, beaucoup plus ancienne, consiste à supposer l'existence d'une intuition intellectuelle. Il est possible de faire remonter l'origine de cette notion à Platon, mais elle fait partie de ces inventions qui, en philosophie, ont servi d'épouvantail à la quasi totalité des traditions contemporaines. Pour être plus exact, disons que l'intuition intellectuelle était mentionnée de temps à autre avant que Kant, ayant décidé de la mettre à mort, dissuade quiconque après lui de déterrer ce cadavre. Or elle pourrait représenter une véritable alternative au dilemme d'explicitation. Cette intuition intellectuelle a en effet la propriété d'être définissable dans le langage philosophique, mais non dans le langage logico-déductif. Platon défend en effet qu'elle permet de contempler les formes des choses, ou les idées, en se plaçant au-delà de la *dianoia*, qui n'est rien d'autre que *l'intelligence analytique*. C'est donc un candidat de premier choix pour opposer une résistance à la prétention algorithmique de réduire tout le champ de la vérité intelligible à la sphère du calcul.

Après ce voyage dans la littérature philosophique existante, je tenterai d'opérer une synthèse personnelle. Je vais en effet admettre que l'intuition de l'infini, dont j'ai montré qu'elle était essentielle au fonctionnement de la conscience, n'est pas susceptible d'être exprimée dans le langage mathématique des Machines de Turing, mais qu'il est toutefois permis de la formaliser dans des machines d'ordre supérieur. Ces machines sont ce que j'appellerai les machines de Turing transfinies. Elles ont cette propriété particulière de pouvoir, tout comme Chuck Norris, compter à l'infini... plusieurs fois. Je proposerai un exemple concret de problème NP-complet où l'être humain et les machines de Turing transfinies sont capables de penser de façon holistique, là où une machine de Turing classique ne parvient pas à dépasser une analyse terme à terme. Je suggérerai que cela peut s'expliquer par cette intuition de l'infini dont j'ai introduit la définition.

9.3 L'Intuition selon Hubert Dreyfus

La définition de l'intuition donnée par Dreyfus apparaît dans un contexte intellectuel très particulier. En 1958, le climat effervescent de la recherche américaine, marquée par le face à face de deux puissances nucléaires, était favorable au développement des technologies de l'information, en ce compris l'intelligence artificielle. Herbert Simon écrivait en effet que l'ordinateur, en

moins d'une décennie, serait (1) en mesure d'asseoir son trône dans le ciel des idées mathématiques en prouvant de nouveaux théorèmes et (2) capable de défier le visage de son créateur en battant les champions humains sur un échiquier.

Dix ans après cette audacieuse prophétie, hélas, la communauté scientifique était forcée d'admettre que les preuves mathématiques sur ordinateur s'apparentaient toujours plus au fac-similé qu'au génie fulgurant de Carl Gauss⁵. Quant aux victoires sur l'échiquier, elles tenaient pour beaucoup à la rapidité de calcul de l'ordinateur en fin de partie et ne se vérifiaient guère en dehors des blitz. Il était donc tentant de penser, dans ces conditions, que la maîtrise d'une tâche cognitive complexe, loin de se réduire, comme les élèves récitant leurs tables de multiplications, à un pur et simple calcul symbolique, faisait intervenir des processus de pensées "intuitifs" qui échappaient aux paradigmes computationnels classiques.

C'est mû de cette présomption qu'un adversaire de Herbert Simon, le professeur Hubert Dreyfus, tentera d'esquisser une théorie de l'apprentissage fondée sur la distinction entre le savoir déclaratif (know-that), facilement manipulable par la machine, et le savoir-faire (know-how) propre à l'humain⁶. Dans cette théorie, exposée en 1986 dans *Mind Over Machine*, l'évolution cognitive dans une discipline passe par cinq stades bien déterminés, allant de la simple compréhension des règles explicites à la maîtrise implicite, et pour ainsi dire *intuitive*, des configurations complexes⁷.

Je vais exposer brièvement ces cinq stades, car non seulement ils représentent une définition fine de ce que *pourrait* être l'intuition humaine ; mais ils ont par ailleurs sensiblement contribué au développement de l'intelligence artificielle. La naissance des réseaux de neurones, dont nous toucherons un

5. En 1978, soit dix ans exactement après qu'Herbert Simon ait formulé sa "prophétie", Haken, Appel et Koch publient un article démontrant qu'un programme d'IBM, de leur conception, permettait de vérifier le théorème des quatre couleurs sur un milliard d'instances. Bien que la correction du programme ait fait l'objet de controverse, et que de nombreuses opérations, dans cette "preuve", restaient de la main de l'homme, ils tentaient ainsi de suggérer la possibilité toute théorique d'un assistant de preuve automatique. Il faudra attendre 1996 pour qu'émerge un premier consensus dans la communauté scientifique, lorsque la preuve sera réduite, par les soins de Robertson, Saunders, Seymour et Thomas, que l'analyse de 633 combinaisons suffit à fournir une preuve concluante.

6. Précisons que Dreyfus ne nie pas la possibilité, pour une machine, d'acquérir un savoir-faire (know-how). Il se contente de dire de ce dernier qu'il ne peut pas, en l'état actuel des connaissances de son époque, être imité informatiquement. C'est d'ailleurs grâce à la critique de Dreyfus que les méthodes de programmation évolueront pour déboucher in fine sur le machine learning et les réseaux de neurones.

7. Cette théorie, précise Dreyfus, a fait l'objet d'une vérification empirique de la part de Patricia Benner. Cette dernière a récolté des données sur des infirmières avec plus ou moins d'expérience (Patricia BENER, *From Novice to Expert : Excellence and Power in Clinical Nursing Practices*, Addison-Wesely, 1984). Ce modèle a si bien influencé le développement ultérieur qu'il est aujourd'hui largement intégré ou débattu (voir notamment PEÑA ADOLFO, *The Dreyfus model of clinical problem-solving skills acquisition : a critical perspective*).

mot dans les chapitres ultérieurs, peut être vue comme la tentative historique par excellence d'implémenter, dans un nouveau paradigme informatique, l'idéal d'apprentissage dépeint par Dreyfus.

- **Stade du Novice** : À ce premier stade, le sujet acquiert un certain nombre d'informations factuelles à propos de la discipline, mais identifie les objets et les actions à réaliser sans référence au contexte. C'est le cas, par exemple, d'un joueur d'échecs qui apprend le déplacement des pièces et utilise la règle du pouce. D'après celle-ci, une dame a deux fois plus de valeur qu'une tour, laquelle, à son tour, en a cinq fois plus qu'un pion — et ce quelles que soient les positions respectives de cette tour ou de ce pion.
- **Stade du Débutant avancé** : Le sujet peut identifier des situations ou des propriétés pertinentes sans être en mesure de décrire les règles d'après lesquelles il en est ainsi. Il utilise cette intuition pour compléter les règles acquises précédemment. Le joueur d'échecs à ce niveau d'apprentissage reconnaîtra une structure de pions avantageuse, ou l'existence de faiblesses dans la position du roi, en se basant sur l'expérience accumulée des situations antérieures. Il continue néanmoins de suivre des règles explicites, hiérarchisées les unes les autres.
- **Stade du Compétent** : Le sujet est capable, face à toute situation, de redéfinir la hiérarchie des opérations à effectuer en vue d'un objectif supérieur. Il n'est donc plus attentif aux règles proprement dites, mais agit d'après le plan ad hoc qu'il s'est fixé. Le joueur d'échecs à ce niveau voit l'échiquier comme un ensemble de tensions conduisant à des objectifs positionnels. Lorsqu'il pense qu'il est temps de mener une attaque sur l'aile Roi, par exemple, il va concéder certaines faiblesses dans sa défense. Il agit sans mobiliser consciemment les règles de déplacement des pièces ou la règle du pouce.
- **Stade du Performant**⁸ : Le sujet acquiert des automatismes. Il n'est plus nécessaire pour lui de les penser rationnellement. En revanche, il doit comparer entre elles les différentes réponses automatiques envisagées, ou se livrer à une analyse pour définir comment conduire sa stratégie d'ensemble. Un joueur d'échecs, se fondant sur sa connaissance de nombreuses situations analogues, aura d'instinct une liste de coups possibles face à une situation donnée, ce qui lui

8. En anglais, Dreyfus utilise le mot « Proficient », qui ne connaît pas vraiment de traduction consensuelle en français. J'ai décidé de lui substituer le terme Performant pour capturer l'idée d'une compétence avancée et profitable. Son étymologie latine renvoie en effet au mot proficiens, participe présent de proficere ("aller de l'avant, faire des progrès, réussir, être profitable ou utile").

permet de jouer très rapidement sans que cela n'entraîne une baisse trop importante de ses performances. Dreyfus ajoute que les joueurs qui voient dans les échecs un jeu analytique restent généralement bloqués au stade 3 de l'apprentissage. Il cite à ce titre le témoignage d'un de ses collègues^[9] :

While students of mathematics and related topics predominate in the population enthousiastics about chess, you are as likely to find a truck driver as a matematician among the world's best player. You are more likely to find a psychologist or a journalist. In a way, I'm glad that my analytical approach to chess stymied my progress, because this help me to see that there is more skill than reasoning^[10].

- **Stade de l'Expert** : L'expert a une connaissance si fine des situations qu'il répond de façon automatique. Il n'a plus vraiment conscience de manipuler des pièces sur un échiquier, mais il est en immersion. Il sent les enjeux bien plus qu'il ne pense objectivement ses actes et ses décisions. S'il se met à penser à la couleur des pièces, ou au fait qu'il est en train de jouer, cela le sort de son état de maîtrise. « Lorsque les choses se déroulent normalement, précise Dreyfus, les experts ne solutionnent pas de problèmes et ne prennent pas de décision ; il font ce qui marche à l'ordinaire^[11] ». En guise d'exemple, il donne l'exemple de Kaplan, un grand maître international qui était toujours en mesure de jouer correctement aux échecs à raison de cinq secondes par coup, bien qu'il dût, dans le même temps, réaliser des opérations arithmétiques. Ce grand maître reconnaît des configurations de jeu comme n'importe qui reconnaîtrait, s'il y était exposé, le visage d'une personne proche. Selon Dreyfus, cette compréhension automatique d'ensemble est rendue possible grâce à une approche cognitive holistique, qui ne cherche pas à réduire le tout à ses parties. En ce sens, il pratique une forme d'arationalité^[12] :

The word rational, derivated from the Latin word ratio, meaning to reckon or calculate, has come to be equivalent to calculative thoughts and so carries with it the connotation of "combining components parts to obtain a whole"; arational behaviour, then, refers to action without conscious analytic decomposition and recombination.

Précisons que les stades de développement ci-dessus n'épousent pas nécessairement le niveau de performance de l'individu, mais bien des types de

9. Mind Over Machine, p.25.

10. HUBERT MIND OVER MACHINE L., STUART DREYFUS E., *Mind over Machine*, Free Press, New York, 1986, p.25.

11. *Ibidem*, p.31.

12. *Ibidem*, p. 36.

fonctionnement : dans l'absolu, un novice de talent peut vaincre un joueur compétent peu inspiré. C'est sans doute la raison pour laquelle un prodige des échecs comme Paul Murphy a pu, dès le plus jeune âge, apprendre les règles en observant son père jouer, mais aussi lui expliquer, à la fin de la partie, qu'il avait commis des erreurs tactiques. Dans certains contextes, le regard du novice peut présenter des avantages sur celui d'une personne plus expérimentée, à plus forte raison lorsque le premier dispose d'une mémoire, d'une adresse ou d'une intelligence fluide hors de l'ordinaire. Loin de voir dans ces cinq stades des paliers de compétences, il faut y discerner des modes de fonctionnement cognitif, dont Dreyfus suppose qu'ils se succèdent les uns à la suite des autres au fur et à mesure que le sujet humain acquiert de l'expérience.

Le fait que Dreyfus illustre ces cinq étapes sur le jeu d'échecs constitue sans doute un quolibet délibéré à l'endroit des prophéties pleines d'optimisme prônées par l'establishment de l'intelligence artificielle¹³. Pour l'anecdote, Dreyfus avait été vaincu sur l'échiquier, dix ans avant la publication de son livre, par le logiciel Mac Hack VI de Richard D. Greenblatt. Cette capitulation, à laquelle avait assisté le fameux Herbert Simon, avait valu l'enthousiasme de ce dernier : « it was a wonderful game — a real cliffhanger between two woodpushers with bursts of insights and fiendish plans ... great moments of drama and disaster that go in such games¹⁴. » Cela n'en a pas moins empêché le philosophe défait de voir, dans le logiciel de ses collègues, rien moins qu'un "novice" doté d'une grande capacité de calcul. Une sorte de Paul Murphy de câbles et de silicium.

Cette idée que l'intuition humaine échappe encore et toujours aux machines pensantes, chez Hubert Dreyfus, tient sa racine dans la conception heideggerienne d'après laquelle l'être humain est un « être au monde ». La propriété essentielle d'un tel être est de sentir le monde à l'intérieur de lui-même, au point qu'on pourrait dire que le monde naît et meurt avec lui. Or puisqu'il en va du destin de tout être humain de disparaître, et de prendre conscience de la contingence de son existence, la question qui se pose à lui est la suivante : pourquoi le monde, ce monde qu'il porte en lui, existe alors qu'il pourrait ne pas être ? La condition humaine est dès lors traversée d'une question métaphysique — une question sans issue — qui pousse l'être humain à tisser, avec son environnement, des réponses relationnelles. Chaque objet est non seulement doté d'une signification logique, mais est aussi l'oc-

13. Signalons que Dreyfus donne tout de même de nombreux autres exemples, notamment celui de la conduite automobile. Pour illustrer le niveau novice, il parlera aussi de la reconnaissance de lettres : un individu qui reconnaît la lettre E en identifiant les trois barres horizontales agit comme un novice. Celui qui la reconnaît par habitude, ou plus précisément, par *pattern recognition*, est un expert (*Mind Over Machine*, p.22). Ironie du sort : ces deux exemples constituent des étapes historiques symboliques dans le développement plus récent des réseaux de neurones.

14. *Oral History of Richard Greenblatt*, interviewed by GARDNER HENDRIE, 2005 Computer History Museum, CHM Reference number : X3131.2005.

casion pour le sujet d'affirmer son existence unique dans le monde. La canne, dans les mains de l'aveugle, est aussitôt intégrée comme une prolongation de ses sens, tandis que pour un homme à la vue impeccable, elle servira plutôt d'adjudant à l'action des muscles moteurs. Elle est en outre porteuse de souvenirs, d'associations d'idées automatiques, d'habitudes, de doutes, d'espoirs. L'outil, entre les mains de l'Homo faber, n'est pas un prêt-à-porter. Il est le dépositaire d'un savoir-faire implicite, qu'en psychologie on nomme l'affordance, et que Dreyfus voit comme le fondement de l'intuition.

Cette affirmation d'après laquelle le savoir-faire (know-how) n'est pas imitable, dans l'état actuel des connaissances, par une machine ; et que cela rend les ordinateurs impropres à l'exécution de certaines tâches dans lesquelles, pourtant, l'être humain peut exceller, va hélas subir un important revers. Lorsque Deep Blue, après plusieurs défaites cinglantes, gagne finalement la rencontre contre Kasparov, force est d'admettre qu'un ordinateur peut atteindre, ou tout du moins imiter, l'expertise tant vantée.

Certes on pourra douter que la machine, même en battant l'être humain aux échecs, agisse par automatisme au même titre que celui-ci. L'automatisme de l'ordinateur, loin de consister en une accumulation d'expériences concrètes, ne serait-il pas un ensemble de calculs bêtes et méchants ? S'il est toutefois une chose indubitable, c'est que le savoir-faire (know-how) introduit par Dreyfus a servi d'inspiration à une nouvelle génération d'informaticiens. Les programmes d'intelligence artificielle, abandonnant la charrue du calcul symbolique, se sont emparé de la machine à vapeur de l'intelligence learning. Tirant profit de ce qu'un logiciel pouvait exploiter une large base de données pour corriger ses erreurs, pour se rendre sensible au contexte, pour repérer des patterns, pour augmenter sa capacité d'interaction avec l'environnement, l'ordinateur s'est tout à coup rendu éligible de cette fameuse intuition qu'on avait voulu lui dérober tout du long. On est alors passé du paradigme symbolique, basé sur le calcul logique, au paradigme connexionniste, basé sur la manipulation de larges libraires de données. Face à cette transition, la philosophie des sciences est mise en demeure, soit de redéfinir ce qu'est l'intuition, et d'explicitier à nouveau en quoi cette intuition distingue l'être humain de l'ordinateur, soit de renoncer à voir, en cette sacro-sainte intuition, une ligne de démarcation, ou tout au moins un point d'achoppement, entre les facultés humaines et les facultés artificielles.

9.3.1 Objections à la position de Dreyfus

Nous en venons ainsi, presque incidemment, à la question capitale : en supposant que l'étude de l'intuition humaine soit d'un quelconque éclairage en épistémologie, est-il possible de décrire ce qu'elle est sans que cette description, aussitôt formulée, puisse, tout au moins théoriquement, être simulée (ou émulée) par un ordinateur ? Lorsque Dreyfus défendait l'existence du savoir-faire (know-how), il soutenait que l'activité de l'esprit ne se

réduisait pas uniquement à un ensemble d'opérations algorithmiques, mais recourait également à la reconnaissance de schémas implicites. La difficulté de cette position consistait néanmoins à décrire ce qu'était le savoir-faire (know-how) en des termes suffisamment explicites pour nous convaincre de son existence, sans ramener, dans un même mouvement, le savoir que nous avons sur ce savoir-faire (know-how) à un simple savoir (know-that) imitable par une machine à calculer. Ce problème est ce que j'ai appelé le *dilemme de l'explicitation*.

Tout se passe donc, en dernier ressort, comme si l'histoire récente de la philosophie de l'intelligence artificielle, en ce qui concerne l'intuition, avait ce double effet paradoxal : d'une part, elle confirme les partisans de l'intelligence artificielle forte dans la conviction qu'il est ultimement possible d'imiter, au moyen d'un mécanisme, toutes les fonctions de la pensée qui échoient traditionnellement au concept d'intuition humaine ; d'autre part, elle confirme aussi les opposants à cette vision dans la conviction que l'intuition ne peut pas être décrite en des termes *purement* explicites (quoique les descriptions puissent lui donner un éclairage).

Précisons cette idée. Les partisans de l'intelligence artificielle forte sont confirmés dans leur optimisme pour cette simple raison historique que, chaque fois qu'une définition de l'intuition est exhibée, et que nous pouvons penser, par exemple, à la sensibilité au contexte, cette dernière définition finit par succomber aux progrès insatiables de la machine. C'est ainsi que la conduite d'un véhicule automatique dans un environnement urbain réel, ou la compréhension fine de la sémantique d'une phrase, tenue autrefois pour irréalisables avec une simple machine, se sont aujourd'hui si bien banalisées qu'il ne viendrait plus à l'idée de personne d'en nier l'automatisation. Il suffira, au lecteur sceptique, de poser la question à ChatGPT¹⁵.

15. J'ai en effet adressé cette interrogation à ChatGPT 4.0, à la date de novembre 2023, et voici le résultat de la conversation : « Question : La compréhension fine de la sémantique d'une phrase constitue-t-elle une tâche irréalisable par la machine ?

Réponse : La compréhension fine de la sémantique d'une phrase est une tâche complexe qui a longtemps été considérée comme irréalisable par les machines. Cependant, les avancées récentes dans le domaine de l'apprentissage automatique, en particulier avec les modèles de traitement de langage naturel basés sur l'apprentissage profond tels que GPT-3, ont considérablement amélioré les performances des machines dans cette tâche.

Ces modèles sont capables d'analyser de grandes quantités de données textuelles et d'apprendre des représentations sémantiques sophistiquées qui leur permettent de comprendre le sens des phrases avec une précision élevée. Les tâches de compréhension de la sémantique, telles que la détection de l'ironie, l'analyse des sentiments ou la compréhension des nuances dans le langage, peuvent désormais être réalisées avec une précision comparable, voire supérieure, à celle des humains.

Cependant, il est important de noter que la compréhension fine de la sémantique reste un défi pour les machines dans certaines situations, en particulier lorsqu'il s'agit de contextes ambigus, de phrases complexes ou de langues avec des structures syntaxiques très différentes. Dans ces cas, les machines peuvent avoir du mal à extraire la signification précise de la phrase et peuvent produire des résultats erronés. Néanmoins, les progrès récents sont encourageants et la compréhension fine de la sémantique par les machines

Dans le même temps, et aussi oxymorique que cela puisse sembler, les opposants aux promesses techniques de l'intelligence artificielle se trouvent confortés dans leur scepticisme. En effet, les échecs sans cesse répétés d'apporter une définition de l'intuition qui résiste à la mécanisation, loin de représenter la preuve que l'humain fonctionne comme une machine, sont, dans leur esprit, autant d'attestations de l'impossibilité de formuler, en des termes mathématiques, un mode de pensée situé au-delà du raisonnement apodictique. N'en va-t-il pas de l'essence de l'intuition que de se soustraire à la logique ? Le cœur de la pensée, pour ainsi dire, a ses raisons que la raison ne connaît point. Pascal dit d'ailleurs à propos des vérités de l'esprit¹⁶ :

Il faut tout d'un coup voir la chose d'un seul regard, et non par progrès de raisonnement, au moins jusqu'à un certain degré. Et ainsi il est rare que les géomètres soient fins, et que les fins soient géomètres ; à cause que les géomètres veulent traiter géométriquement les choses fines, et se rendent ridicules, voulant commencer par les définitions, et ensuite par les principes, ce qui n'est pas la manière d'agir en cette sorte de raisonnement.

9.3.2 Sortir du statu quo

Une façon simple et élégante de se départir de cette contradiction de principe, sans pour autant tomber de Charybde en Scylla, consisterait à définir l'intuition au sein d'un formalisme qui, bien que parfaitement explicite, serait hors de portée de toute machine possible. Une telle stratégie peut se décliner de deux façons évidentes :

1. Le concept d'intuition ne peut pas être explicité à l'aide du langage ;
2. Le concept d'intuition ne peut être explicité qu'au travers d'un langage non-mathématique ;
3. Le concept d'intuition peut être explicité dans un langage mathématique sans être pour autant réalisable par une machine ;

La première posture est celle du mystérianisme ; la seconde est celle du Platonisme. Elles feront l'objet des deux sections qui suivent. La troisième posture, pour finir, est la réponse que je vais tenter de construire après avoir brièvement montré les difficultés des deux précédentes démarches. Elle se présentera donc comme la voie du juste milieu. C'est l'élément de synthèse réunissant habilement la thèse et l'antithèse.

Avant d'aller plus loin, je me dois de faire une observation. À l'instar du balayeur qui, discrètement, glisse la poussière incommode sous le tapis, le dilemme, tel que je l'ai écrit, semble passer à la trappe l'idée d'abandonner tout simplement le concept d'intuition, et d'affirmer — pourquoi pas ?

continue de s'améliorer. »

16. PASCAL Blaise, *Pensées*, 305.

que l'esprit ne serait rien d'autre qu'un ordinateur sophistiqué. Je pense d'ailleurs que, dans la doxa du monde contemporain, il ne s'agisse pas de l'opinion la moins répandue. Dans la mesure, toutefois, où le rôle de cet ouvrage est de construire un argument permettant de définir, si elles existent, les limitations théoriques intrinsèques du programme de l'intelligence artificielle, le dilemme que nous avons exposé semble incontournable. Ce n'est qu'en prenant au sérieux l'idée d'après laquelle il est possible d'explicitier l'intuition humaine que l'on verra bien si, oui ou non, cette intuition résiste aux machines de Turing.

On pourrait aussi arguer, de façon analogue, du fait que ce n'est pas l'intuition qui joue un rôle clef face au pouvoir de calcul des machines, mais bien la mémoire, la corporéité ou l'intentionnalité. À procéder de cette manière, on n'en échappe pas moins au problème central. Quelle que soit en effet la faculté *X* exhibée pour faire obstacle à la mécanisation, ou le nom qu'on souhaitera donner à cette faculté, cette dernière n'aura pas d'autre choix que d'être soit définie en des termes calculables et effectivement exécutables à l'aide d'un algorithme, soit en des termes calculables non exécutables par un algorithme, soit en des termes non calculables, soit de ne n'être pas définie du tout — ou toutefois pas de façon explicite.

9.4 Première solution : le Mystérianisme

Dans le précédent chapitre, j'ai parlé du dilemme de l'intuition en cachant sous le tapis cette réponse : que l'intuition ne soit pas, conceptuellement parlant, définissable en des termes objectifs et scientifiques. Cette alternative est attrayante, à plus forte raison si l'on admet que l'intuition présente quelque chose d'irreprésentable, liée étroitement à la conscience humaine. Ce lien entre conscience et intuition est motivé par le fait que l'intuition s'impose à nous comme quelque chose dont on est conscient, mais qu'on ne sait pas toujours expliquer. L'absence d'explication tient d'une part du fait que l'on ne peut pas tracer les processus inconscients qui ont donné naissance à cette intuition. Elle tient d'autre part du fait qu'on ne peut pas expliquer la conscience elle-même au moyen de mécanismes physiologiques clairement définis.

Je ne voudrais pas donner ici l'impression de remplacer un mystère par un autre mystère, plus abscons encore que le précédent. Je me contente de dire qu'il existe des choses mystérieuses au sein de l'esprit humain. La question n'est pas tant de savoir si l'intuition est comparable à la conscience, que de se demander si l'intuition ne ferait pas partie de ces mystères que l'on vit, mais qu'on ne sait pas expliquer. Alan Turing lui-même écrivait : « Je ne veux pas donner l'impression de penser qu'il n'y ait, à propos de la conscience, rien de mystérieux. Il y a, par exemple, quelque chose de l'ordre

d'un paradoxe lorsque l'on tente de la localiser¹⁷. » N'éprouve-t-on pas, de façon analogue, une difficulté à localiser l'intuition dans le cerveau humain ?

Dans cette section, j'aimerais discuter le courant *mystérianiste*, qui défend que le problème de la conscience n'est pas susceptible d'être résolu par un esprit humain. Pour ce faire, je commencerai par faire remonter le mystérianisme, en tant que courant philosophique, au discours inaugural d'Emil-Dubois Reymond. Je tenterai ensuite d'esquisser une position plus contemporaine.

9.4.1 Ignoramus et ignorabimus

En 1880, lors d'un discours à l'Académie royale des sciences de Prusse, Emil du Bois-Reymond a fait la liste de sept problèmes qui, d'après lui, ne pourraient jamais recevoir d'explication scientifique. Au rang de ces monstres immortels de la pensée humaine, se trouvent la nature de la matière et de la force, l'origine du mouvement, l'origine de la vie, l'apparente téléologie (ou finalité) de la nature, l'origine de simples sensations, le libre arbitre et l'origine de l'intelligence elle-même. Par ce geste inaugural, Du Bois-Reymond introduisait *l'ignoramus et ignorabimus* : nous ne savons pas et ne saurons pas.

Il faut reconnaître que l'histoire des idées semble pencher en faveur de Du Bois-Reymond. N'est-il pas vrai que certains problèmes fondamentaux, dont la liste pourrait être discutée, n'ont pas reçu de progrès substantiels depuis leur formulation dans les limbes de la philosophie grecque ? Que certaines questions ont le pouvoir de plonger l'esprit dans une torpeur métaphysique ? Il reste néanmoins curieux de constater que, quand bien même la réponse à ces questions n'existerait pas, nous sommes tout de même capables de les formuler. Elles ne sont donc pas impensables. On est dès lors en droit de se demander comment il se peut qu'une question, pour laquelle il n'existe aucune réponse, puisse tout de même être posée ? Ne serait-il pas plus logique de penser que si la question était vraiment impossible, il nous serait tout aussi impossible de nous la représenter ?

Du Bois-Raymond se comporte comme ces athlètes qui prennent leur élan avant de s'arrêter devant le précipice. Pour pousser jusqu'au bout sa démarche, n'aurait-il pas fallu demander, de façon à la fois plus modeste et plus radicale : peut-on seulement savoir s'il existe ou non une réponse à tous les problèmes ? Et peut-on savoir s'il existe une réponse à la question de savoir s'il existe une réponse ? Toujours est-il que l'état de perplexité était encore si patent, cinquante ans après l'allocution de Du Bois-Reymond, que le mathématicien Hilbert prendra l'initiative de couper le nœud gordien :

17. « I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. », dans TURING A., *Computing Machinery and Intelligence*, p.12. La traduction est de moi.

« Il n'y a pas d'ignorabimus en sciences naturelles. Nous devons savoir, nous saurons. »

Il est important de distinguer le mystérianisme de deux autres positions qui s'y apparentent, sans être *stricto sensu* équivalentes : l'Aporétisme et le Mysticisme. Ces deux dernières attitudes ont ceci de commun qu'elles reconnaissent qu'il existe des questions sans réponses. Elles admettent donc ce fait étonnant que nous pouvons formuler une question à propos de quelque chose que nous ignorons ; elles expliquent toutefois ce fait de façons bien différentes :

1. **L'Aporétisme** : les questions que nous posons sont formulées en des termes contradictoires. Si elles n'admettent pas de réponse, c'est parce qu'elles ne constituent pas un vrai problème. Nous sommes toutefois pas assez agile, intellectuellement, pour repérer la contradiction cachée derrière toutes nos vaines questions. Si notre langage était épuré, il n'y aurait plus de faux problèmes.
2. **Mysticisme** : les questions que nous posons, bien que parfaitement cohérentes, n'admettent parfois aucune réponse scientifique, aucune réponse philosophique, voire aucune réponse tout court. La raison en est que la réalité n'est pas totalement intelligible. Nous ne pouvons appréhender les mystères de l'Univers qu'au moyen de facultés non rationnelles.
3. **Mystérianisme** : Les questions que nous posons admettent des réponses, mais notre intelligence humaine ne nous met pas en position de les comprendre. Les mystères apparaissent ainsi comme une mise en lumière des limites intrinsèques de notre langage.

Le Mystérianisme ne se contente donc pas de dire qu'il existe des questions sans réponse. Il fait le pari que les réponses, bien qu'elles existent, sont hors de portée de notre esprit. Pour illustrer cette position, et la connecter à notre dilemme de l'intuition, je voudrais me pencher sur un texte du philosophe britannique Colin McGinn, intitulé *Can We Solve the Mind-Body problem ?* Bien que ce texte concerne la conscience, et non l'intuition, je m'intéresserai surtout à la structure argumentative.

9.4.2 La posture de McGinn

Dans un article intitulé *Can We Solve the Mind-Body problem ?*, écrit en 1989, Colin McGinn cite Julian Sorell Huxley non sans une certaine éloquence¹⁸ : « Le fait que quelque chose d'aussi remarquable qu'un état de conscience se produise par suite de l'initiation d'un tissu nerveux est tout aussi inexplicable que l'apparition du Djin, après qu'Aladdin eût frotté

18. Julian Sorell Huxley (1887 – 1975), biologiste évolutionnaire et eugéniste, ne doit pas être confondu avec Aldous Huxley (1894 – 1963), l'auteur du *Brave New World*.

la lampe magique¹⁹. » Par ces quelques mots, c'est toute une tradition explicative de la conscience qu'il entend remettre en doute.

Pour asseoir ses vues, il précise pour commencer qu'il existe deux ordres possibles d'explications. La première, qu'il nomme l'approche *constructive*, consiste à chercher un ensemble de propriétés qui soient satisfaites, de façon concomitante, à la fois au sein des états mentaux et au sein du cerveau (ou dans le corps). La seconde approche, dite *dualiste*, et historiquement dominante à ses yeux, postule l'existence d'entités surnaturelles ou d'intervention divine. Tout se passe ainsi comme si l'esprit humain était placé entre deux solutions, pour le moins, aussi mystérieuses que le phénomène à expliquer : un cartésianisme de substance ou une harmonie préétablie à la mode de Leibniz.

Il attire ensuite l'attention du lecteur sur le fait qu'il pourrait exister une clôture cognitive au sein de l'esprit humain. Cela signifierait qu'il existerait des propriétés qui, bien que scientifiquement exactes et explicables, ne puissent pas être comprises par notre espèce. La clôture cognitive, en toute rigueur, ne doit pas être confondue avec la clôture perceptive qui, quant à elle, désigne la capacité à appréhender un phénomène au moyen de nos sens. À titre d'exemple, il ne fait pas de doute que nos perceptions soient closes vis-à-vis de l'atome, qui ne se laisse pas observer sous le microscope. Cela ne signifie pas pour autant que l'atome ou les fréquences invisibles ne puissent pas être conçues dans l'esprit humain. Tout au contraire, l'atome, pris dans sa signification première — à savoir : l'unité explicative élémentaire — semble si essentielle à l'édification des théories humaines qu'il n'a pas fallu attendre la physique contemporaine pour être exposé à son concept.

Il est à noter, ajoute McGinn, que, pour un esprit humain classique, les idées devraient se former à partir des impressions, si bien que tout ce qui est clos perceptivement devrait être clos cognitivement. Une idée semblable est d'ailleurs attribuée à Locke, qui « pense que nos idées de la matière sont étroitement contraintes par nos perceptions et conclut ainsi que la vraie science de la matière est éternellement au-delà de nous — et qu'on ne pourra jamais se débarrasser de notre perplexité vis-à-vis de la question de savoir ce qu'est la solidité²⁰ » Mais bien que l'esprit humain ne soit sans doute pas humain en ce sens restrictif, il doit se conformer à certains principes liant les impressions aux idées, et les idées entre elles, si bien que, l'esprit n'étant pas en mesure d'appréhender tous les phénomènes sensoriels, il est raisonnable

19. Je me suis permis de traduire depuis l'anglais : "How it is that anything so remarkable as a state of consciousness comes about as a result of initiating nerve tissue, is just as unaccountable as the appearance of the Djinn, where Aladdin rubbed his lamp in the story... (Julian Huxley)"

20. Cette citation est une traduction de ma part. La phrase originale peut être trouvée dans MCGINN COLIN, *Can We Solve the Mind-Body problem*, 1989. Dans une note de bas de page, McGinn précise que cette idée de Locke vient de son Essai sur l'Entendement humain, livre II, chapitre IV.

de penser qu'il ne puisse pas non plus concevoir toutes les idées possibles.

La question qui s'adresse dès lors à nous est la suivante : la conscience est-elle cognitivement close pour l'espèce humaine ? Une telle position, aux yeux de McGinn, n'induit pas de mysticisme. Le phénomène de la conscience proprement dit n'est pas en dehors du champ scientifique — le philosophe semble d'ailleurs insister là-dessus avec beaucoup de véhémence. Il en va finalement de la conscience comme de la vie. À moins de supposer l'intervention du doigt de Dieu, il est raisonnable de penser que la vie se soit formée sur la Terre à partir de la matière organique connue. Néanmoins, il se pourrait que l'explication soit hors de portée de l'esprit humain, au même titre que la théorie de la relativité est hors de portée de l'intelligence d'un enfant de cinq ans.

Bien entendu, un argument en faveur de cette thèse serait de souligner les échecs que la science a enregistrés sur cette difficile question. Une telle stratégie est l'analogie de l'argument méta-inductif pessimiste que l'on mobilise parfois contre le réalisme métaphysique, stipulant qu'il existe un monde en dehors de nos sens.^[21] Dans la mesure, toutefois, où de nombreux phénomènes scientifiques ont pu être pensés insolubles avant d'être finalement compris de l'esprit humain, il ne s'agit en aucun cas d'un argument significatif.

McGinn préfère alors opter pour une démarche qui n'est pas sans rappeler la distinction que nous avons opérée, dans l'introduction, entre la conscience noématique, ou objective, et la conscience noétique, ou subjective. Les deux voies d'approche, pour comprendre la conscience, d'après le philosophe, sont en effet l'introspection et l'observation scientifique. Bien que l'introspection donne un accès "immédiat" au contenu de la conscience, elle s'exerce dans un champ d'intelligibilité pour ainsi dire indifférent au corps. En effet, il se pourrait qu'aucun corps n'existe, et l'introspection, par principe, ne pourrait pas moins théoriquement nous donner accès à quelque chose. Elle pourrait de même nous faire apercevoir des choses qui n'ont aucune contrepartie dans ce que nous nommons communément la réalité, à l'instar du rêveur. D'autre part, l'observation scientifique nous décrit le fonctionnement des corps, tels du moins que nous nous les représentons dans nos théories, mais sans fournir les contenus de conscience qui sont mobilisés. Pour le dire autrement, la conscience n'est pas un élément observable au sein du cerveau (à supposer qu'elle soit dans le cerveau), comme le serait par exemple un objet, une couleur, une texture, etc.

21. Pour une revue de cet argument, voir par exemple STANFORD HOWDYSHELL, *On Realism and the Pessimistic Meta-Induction*, Kent State University, Critical Reflections, Essai on Significance and Critical Reflections.

9.4.3 Principales objections

Bien que l'idée générale de McGinn s'inscrive dans la même lignée que les chiasmes internalistes et externalistes qui ont été décrits dans l'introduction, certaines critiques pourraient être émises. Pour commencer, il est naïf de penser que l'introspection donnerait un accès direct à la conscience. Il serait plus plausible de dire qu'elle donne certes accès à un contenu, mais cet accès lui-même est conditionné par toutes sortes d'éléments extérieurs à l'introspection proprement dite. Je ne parle pas du fait d'avoir un cerveau, car ce serait commettre là une pétition de principe. Dans l'état d'introspection, le penseur en quête d'une solution au *Mind-Body problem* ne peut pas affirmer ou infirmer l'idée que, sans son cerveau, l'introspection serait impossible ; car ce dernier est une réalité existant au-delà des sens. Toujours est-il que l'introspection se réalise au moyen de mots du langage, de catégories conceptuelles et sensorielles, qui ne semblent pas être le produit de l'individu lui-même. Il y a donc, au sein de l'introspection, référence à quelque chose d'autre que notre propre conscience. Cette altérité fondamentale est ce qui permet de douter de l'individualisme méthodologique : il se pourrait que l'introspection ne soit pas une faculté dotée en propre par l'individu humain ; mais bien plutôt le résultat d'un ensemble de construits socialement déterminés. Dans cette optique, toute introspection est médiatisée par le langage, les concepts, et, plus généralement, par notre culture.

Une autre critique concerne, cette fois, l'idée que l'observation scientifique donnerait accès au corps. Dans la mesure, en effet, où cette observation se réalise toujours à partir d'un esprit humain, il n'est pas forcément clair de savoir en quoi elle ne constituerait pas, à son tour, une sorte d'introspection, bien que médiatisée différemment. Une posture scientifique véritablement rigoureuse demanderait de sortir de notre subjectivité et de s'exprimer, pour ainsi dire, du point de vue de Dieu. Dans la pratique de l'espèce humaine, et c'est bien le point de départ abordé par McGinn, cela est impossible. La distinction même entre introspection et observation semble dès lors compromise. C'est pourquoi nous préférons parler, dans l'introduction, de noème et de noèse, d'objectivité et de subjectivité. L'introspection, c'est de l'observation qui s'applique aux états de pensée. L'observation, c'est de l'introspection qui suppose que les idées auxquelles cette introspection réfère existent dans le monde.

Cela étant posé, il est vrai que le cerveau peut être conçu comme un objet du monde. Il suffit de faire abstraction, lorsque nous nous représentons le cerveau, de l'acte de perception lui-même, pour ne garder que l'idée d'un cerveau existant indépendamment de nous. Il est d'ailleurs remarquable que l'être humain ait la capacité d'envisager que les choses qui existent dans son esprit existent également en dehors de lui — et inversement. Cette frontière conceptuelle entre l'objet en soi et l'objet pour nous est probablement une des propriétés les plus essentielles de l'esprit. Sans celle-ci, nul doute que la

plupart des problèmes métaphysiques n'existeraient pas ! C'est cette frontière qui fait que, lorsque j'envisage le cerveau comme un objet en soi, je fais abstraction de ma perception. Il n'y a plus, dès lors, rien, dans ce cerveau, qui soit de l'ordre d'une perception. En ce sens précis, on peut dire que l'observation scientifique ne trouvera jamais, dans les objets ainsi considérés, rien qui se rapporte à la conscience.

Nous pensons donc que l'antinomie du problème Corps-Esprit vient du fait qu'il y a contradiction dans les termes. L'observation, par définition, se dépouille de l'objet pour soi de la perception ; et l'introspection, de son côté, fait abstraction de l'objet en soi de l'observation. Si bien que le véritable mystère, si mystère il y a, consiste à savoir pourquoi nous sommes en mesure de distinguer l'en soi du pour soi. Il va sans dire que certains ingrédients peuvent, d'ores et déjà, être soulignés. Tout d'abord, il semble que cette capacité à opérer la distinction soit co-dépendante de la capacité à être conscient de soi comme un "je pense" cartésien. Si, en effet, il n'y avait d'en soi et de pour soi, il n'y aurait pas de différence entre ce que nous pensons du monde, en tant que conscience, et ce que nous pensons qui existe au-delà de notre pensée. Il ne serait donc pas permis, ni de se penser dans le monde, ni de penser le monde en nous. La philosophie se voit confier la tâche quasi sacrée d'entreprendre l'identification des conditions qui rendent cette opération de pensée *possible*. Une telle démarche sera effectuée dans la dernière partie de cette thèse, qui entreprendra de reconstruire le concept de « Je » phénoménologique. Dans l'immédiat, j'espère avoir donné, au travers de ces quelques remarques, une révision intéressante à la posture de McGinn.

9.4.4 L'Intuition est-elle un mystère ?

Avant de finir, il est légitime de se demander en quoi le mystérianisme, qui concerne la conscience humaine, est-il porteur d'une réponse à la question de l'intuition ? La question consistant à résoudre le problème Corps-Esprit n'est, en substance, pas tout à fait équivalente à celle consistant à dresser une démarcation entre la pensée de l'homme et de la machine. Cette dernière question, en tout cas, bien qu'elle supporte la question métaphysique, peut être comprise en un sens plus fonctionnaliste : indépendamment de toute explication sur la conscience proprement dite, quelles sont les fonctions de pensée inaccessibles à la machine dans les paradigmes actuels ? Il se trouve que certaines élaborations récentes du mystérianisme apportent un éclairage intéressant.

Dans un article intitulé *Level-headed mysterianism and artificial experience*, Prinz propose de définir le mystérianisme de la conscience au moyen d'une analogie : combien de cuillères de cyanure faut-il pour s'intoxiquer²² ? Nous savons de source sûre qu'un milligramme ne suffit pas. Certes, il faudra

22. J. PRINZ, *Level-headed mysterianism and artificial experience*, Journal of Consciousness Studies 10 (4-5) :111-132, 2003

au moins cette quantité-là, mais ça ne suffira pas. C'est donc une condition nécessaire, mais non suffisante. Inversement, et à moins d'être la résurrection diabolique de Mithridate, nous savons qu'une tonne est amplement suffisante, mais qu'il n'est pas nécessaire d'en ingérer autant pour passer *ad patres*. Il s'agit donc, cette fois, d'une condition suffisante, mais non nécessaire.

À partir de cet exemple, il est facile de comprendre l'idée de Prinz : la conscience ne serait-elle pas ce dosage inconnu ? Ce mystérianisme revisité consiste à dire que, quelle que soit la méthode utilisée, il est impossible de déterminer l'ensemble des conditions nécessaires et suffisantes à la présence de la conscience dans un organisme donné. Cela signifie que, chaque fois qu'une propriété est envisagée, soit c'est une condition suffisante (il est suffisant, pour avoir une conscience, d'être un Je phénoménal), soit c'est une condition nécessaire (il est nécessaire d'être un organisme vivant, d'avoir de la mémoire, des perceptions...), mais jamais les deux en même temps.

Outre le soulagement que cette stratégie induit, comme lorsqu'un noyé revient soudainement à la vie, je propose l'idée suivante : le concept d'intuition contient au moins un des éléments X de la conscience qui la rend mystérieuse. Autrement dit, ce X est à l'intersection de l'intuition et de la conscience. La posture mystérianiste consistera à dire qu'il est impossible de donner les conditions nécessaires et suffisantes pour avoir X, si bien qu'on ne peut pas non plus donner les conditions pour avoir l'intuition ou pour avoir la conscience.

Dès lors, et on en vient à la question que nous nous posons, chaque fois que l'intuition capture ce X, elle ne pourrait pas être explicitée à l'aide du langage. Ce que je veux dire par là, c'est qu'il y a deux possibilités. Soit l'intuition est définie de telle sorte à ne pas exprimer la propriété X qui fait l'objet de tous les mystères, soit elle exprime ce X. Dans le premier cas de figure, on peut lui donner des conditions nécessaires et suffisantes. Pour peu que ces conditions soient des processus logiques imitables par une machine de Turing, on finira par construire un ordinateur capable d'intuition. Dans l'instant que l'on veut dire quelque chose à propos de X, il ne sera pas possible de donner une définition nécessaire et suffisante de l'intuition.

Notons que même si les machines se montraient capables d'imiter tendanciellement toutes les fonctions usuellement attribuées à l'intelligence humaine, elles ne convergeraient pas forcément vers le X mystérieux. Il faudra certes admettre que toutes les fonctions de cette machine sont nécessaires à l'intelligence humaine, mais cela ne rendra pas ces conditions suffisantes. Ce cas de figure est celui qu'on obtient, en mathématique, dans le cas de suites convergentes qui ne sont pas définies en leur point de convergence.

Illustrons cela géométriquement en supposant que l'on demande à un enfant de construire un cercle en traçant sur une feuille des segments de même longueur. Considérons en outre que l'enfant, bien qu'il connaisse parfaitement la définition des termes figurant dans la question, n'a aucune idée

quant à sa réponse. Cet enfant est dans une situation analogue à l'enfant de cinq ans, décrit par McGinn, qui essaierait de conceptualiser la théorie de la relativité.

L'enfant peut néanmoins se rendre compte que, s'il utilise trois segments, il obtiendra un triangle. Que s'il en utilise quatre, il aura un carré ; et que le carré est déjà plus proche d'un cercle. Continuant ainsi l'impossible quadrature, il observera qu'un polygone régulier à n côtés est plus proche du cercle qu'un polygone régulier disposant de moins de côtés. Au fur et à mesure que n augmente, et pour peu qu'il soit un peu futé, il peut même prouver que la somme des longueurs des côtés du polygone, divisée par la distance du centre à l'un quelconque des sommets, se rapproche de π (mais en étant chaque fois plus grand). Notons ce rapport P . Il est donc nécessaire, pour être un cercle, que P soit au moins plus petit que celui que l'on trouve dans un polygone régulier de n côtés. Le problème vient de ce que, quelle que soit la taille de n , P ne sera jamais égal à π . Nous avons donc généré une infinité de conditions nécessaires, mais insuffisantes.

Nous arrivons à la conclusion que, pour un être ne connaissant qu'un nombre strictement fini de segments, le fait d'être un cercle est inintelligible. Ce n'est pas seulement une énigme mathématique, mais c'est également un au-delà de la pensée. Cet au-delà, néanmoins, est présent dans son esprit. Il existe, pour ainsi dire, tendanciuellement — ou asymptotiquement. Le cercle, loin d'être seulement un mystère dans la tête de notre enfant, comme c'était le cas avant qu'il n'explore la question, est devenu quelque chose de proprement mystique. Le cercle était d'ailleurs pour les pythagoriciens, ces enfants de la pensée philosophique, la forme parfaite. On pourrait donc dire, d'après cette posture, que le X mystérieux qui fait tout le sel de la conscience et de l'intuition humaine, cet Omphalos métaphysique, ce centre névralgique vers lequel toutes les questions philosophiques semblent converger, c'est quelque chose dont on peut se faire une idée tangentiellement, mais qu'on ne saura jamais expliciter effectivement.

9.5 Deuxième solution : l'Intuition intellectuelle

Jusqu'à présent, nous avons parlé du dilemme de l'explicitation comme s'il ne comportait que deux termes antagonistes : soit l'intuition peut être explicitée et elle est donc réductible à un ensemble d'opérations intelligibles, soit elle ne le peut pas. Dans cette section, j'aimerais suggérer une tierce voie : exprimer l'intuition au sein d'une théorie qui ne soit pas mathématique, et donc *de facto*, pas réductible à une déduction automatique — mais qui n'en resterait pas moins explicite dans le langage. C'est une attitude qu'il est possible de faire remonter à Platon. On lui a donné le nom, oxymorique s'il en est, *d'intuition intellectuelle*. Je vais en donner la définition en me penchant sur le texte grec d'origine, après quoi je m'intéresserai aux critiques que

Kant en a faites. Je terminerai en définissant une stratégie pour résoudre le dilemme de l'explicitation.

9.5.1 La Contemplation des Idées

Platon pose les jalons de l'intuition intellectuelle dans un passage de la *République* que j'ai déjà étudié au début de cette thèse : l'analogie de la Ligne. Au risque de me redire, je pense qu'il est important de montrer en quoi ce texte est révélateur de l'idée qu'il existe, chez les Anciens, un ordre de vérité différent de la simple logique déductive. Ce sera également l'occasion de comprendre pourquoi la critique de Kant, en ramenant les vérités mathématiques au niveau des jugements synthétiques *a priori* de l'Entendement, a rendu cette voie impraticable²³.

Dans L'Analogie de la Ligne, on voit Socrate s'entretenir avec Glaucon sur la nécessité que la cité soit gouvernée par les philosophes et sur les vertus de ces derniers. Il compare l'organisation politique à un navire dont le capitaine serait certes grand et fort, mais à moitié borgne, à moitié sourd et peu vivace. Les matelots se disputent la barre, endorment la vigilance de leur chef avec de la mandragore ; et s'il arrive que l'un d'entre eux s'empare du gouvernail, les autres le tuent et le jettent par-dessus bord. Dans ce contexte, le sage, celui qui connaît réellement l'art de naviguer, est mis au ban de la société. Il est pourtant le seul capable de « voir », mais sa bonne nature, si elle n'est pas corrompue, paraît aux autres comme la plus excentrique.

Le tempérament d'un sage, précise Socrate, est comme une plante croissant dans un climat étranger. Elle a besoin de plus de soins pour ne pas dépérir. Et cette plante qui est plus grande peut porter les meilleurs fruits, comme elle peut donner naissance au pire d'entre eux. Dans le bien comme dans le mal, elle ne sera en aucun cas victime de médiocrité. C'est pourquoi la vertu qui lui est la plus nécessaire est celle de concevoir l'idée du bien. Cette idée, cependant, ne relève pas du monde sensible. Il ne s'agit pas de la multitude des choses bonnes, considérées séparément, pas plus du moins que le beau ne serait la multitude des beaux tableaux. Elle est bien plutôt comme le soleil ; source de lumière, rien ne peut être vu clairement sans elle. Ôtez-la et notre philosophe est tout à coup aveugle à la vérité. Mais tout comme le soleil, elle ne peut pas être dévisagée en face, les yeux grands ouverts. La seule façon de s'y hisser à cette idée du bien est de partir d'hypothèses et, en raisonnant par étape, de s'élever jusqu'à l'idée située au-delà de toute hypothèse, que l'on nomme l'anhypothétique.

À ce titre, la démarche du philosophe peut être apparentée à celle du géomètre. Les disciples d'Euclide parlent certes de triangles, de carrés et de

23. Cette redite que je me permets de faire ici, en reprenant Platon et Kant comme point d'ancrage, est également justifiée par le fait que j'essaie de donner un maximum d'autosuffisance aux différentes parties de ma thèse. Mon ambition est en effet de la faire publier sous forme de plusieurs essais certes complémentaires, mais individuels.

diagonales, mais ce qui les importe ne sont pas les formes sensibles dessinées dans le sable pour les aider à raisonner : ce sont les idées mathématiques en elles-mêmes. Il existe cependant, chez Platon tout comme c'était le cas de Pascal, une différence essentielle entre les géomètres et l'intelligence la plus haute. Les géomètres, en effet, ne voient que les principes et ils s'appuient sur des images pour raisonner sur ceux-ci. Ils procèdent par esprit d'analyse et de déduction. Le sage, en revanche, lorsqu'il envisage l'idée du bien, n'a pas de copies sensibles à sa disposition. Il ne peut pas dessiner le bien dans le sable et en déduire quelque chose de l'idée essentielle. Il doit formuler des hypothèses situées au-delà de la représentation visible. La contemplation de l'idée du bien n'est donc pas le fruit d'une déduction ou d'une analyse, mais d'une contemplation, d'une synthèse, ou encore, d'une intuition intellectuelle.

La vérité, d'après Socrate, peut donc être représentée le long d'une ligne divisée en quatre segments de même longueur. Les deux premiers segments représentent les deux niveaux de réalité du monde sensible. Les deux derniers segments représentent quant à eux les deux niveaux de réalité du monde intelligible : celui des raisonnements logico-déductifs (*Diánoia*) et celui des Formes (*noèton*). J'aimerais rappeler ci-dessous comment Socrate oppose ces raisonnements logico-déductifs, qui relèvent du simple calcul *Diánoia*, à la faculté de contempler intuitivement les vérités supérieures (*Noèsis*). Il y a en effet, au détour de ces pages, une affirmation inaugurale : la faculté la plus élevée de l'homme ne consisterait pas à calculer à la façon des machines d'aujourd'hui, mais à contempler le savoir, à le sentir, à s'en imprégner avec extase et fascination, à la façon dont Galilée cherchait à regarder le Soleil :

Noèton <i>Epistémé</i> (science)		Horatón <i>Doxa</i> (opinion)	
<i>Noèsis</i> (intuitif)	<i>Diánoia</i> (discursif)	<i>Pistis</i> (croyance)	<i>Eikasia</i> (imagination)
↓	↓	↓	↓
porte sur l'anhypothétique	porte sur l'hypothétique	porte sur les objets sensibles	porte sur leurs images

— **Dianoia** : Traduit par « connaissance discursive » (Chambry, Bacou), « discursion » (Robin), « raison » (Dixsaut, Karsenti/Prélorentzos) ou simplement « pensée » (Pachet, Cazeaux, Leroux), la signification de ce terme recouvre en grande partie celui de *noèsis*, que nous verrons ci-dessous, mais diffère de lui par l'adjonction de la particule « dia ». Quelle est sa signification ? Dans les usages attestés, elle peut induire une idée (1) de séparation, au sens d'une distribution ou d'une division ; (2) de pénétration ; (3) de supériorité ou encore (4) d'achèvement. Les acceptions les plus pertinentes ici sont probablement celles de la séparation ou de la division. La fa-

culté que Socrate a en vue, concernant ce segment, est proprement discursive. Il s'agit de structurer le contenu de la sensibilité de façon à en extraire des principes intelligibles. À la façon du géomètre qui décompose les figures géométriques en points et en segments, il est tout d'abord requis de diviser les images sensibles ou les mots du langage en autant d'éléments d'analyse que nécessaire. En posant un certain nombre d'hypothèses, l'on s'élève ensuite vers certains principes, ou certaines thèses, de validité universelle. L'archétype de ce raisonnement est celui des mathématiques, dans leur dimension formelle et apodictique. Le discours associé est logico-déductif. Notons, pour compléter cette définition, que la particule « dia » peut aussi se référer au fait que la *dianoia* occupe l'entre-deux sur la droite de l'être. D'un côté en effet se tient l'intelligible pur, de l'autre côté se tiennent les images. Elle peut donc être conçue comme la faculté qui sert de médiation entre un savoir des images sensibles et un savoir des Formes. Elle s'exerce de plus au travers d'un dialogue que l'âme se tient à elle-même.

- **Noësis** : Ce terme a été rendu par « intelligence » (Chambry, Baccou, Dixsaut, Pachet, Karsenti/Prélorentzos), « intellection » (Robin, Leroux) ou encore « activité spirituelle » (Cazeaux). Il est formé du verbe *noein* au moyen du suffixe *-sis*. D'usage plus rare que *nous*, il a cet avantage de désigner l'intelligence en tant que faculté, et non dans ses autres sens. C'est donc bien d'une certaine capacité à appréhender l'intelligible dont il est question ici. Contrairement à la *dianoia*, il ne s'agit pas tant de partir d'hypothèses et de procéder par voie déductive, que de questionner le statut ontologique et épistémologique de ces hypothèses. La démarche consiste à chercher le premier principe, anhypothétique, grâce auquel il est permis de concevoir la véritable nature des choses. Or comme le principe du Bien, on l'a dit, ne tombe pas sous le raisonnement discursif ordinaire, cette faculté d'atteindre à l'intelligible pur ne peut être que de nature intuitive. Il s'agit à proprement parler de cette partie de la raison qui, lorsqu'elle s'émancipe des sens, peut être conduite à la révélation et à la contemplation de l'Idée du Bien. C'est pourquoi nous la traduirons par raison intuitive ou encore, contemplative. Socrate dit de celui qui en use « [qu'] il a recours à la construction d'hypothèses sans les considérer comme des principes, mais pour ce qu'elles sont, des hypothèses, c'est-à-dire des points d'appui et des tremplins pour s'élancer jusqu'à ce qui est anhypothétique, jusqu'au principe du tout²⁴ » La philosophie moderne a dépouillé ce qu'il y avait de poétique dans la pensée de Platon, sans toutefois perdre la notion « d'intuition » ou « d'évidence ».

24. *République*, 511b.

Cette notion de contemplation chez Platon, quoique essentielle à sa pensée, est sans doute un héritage du pythagorisme. Loin de se confiner à un enseignement théorique d'ordre purement rationnel ou symbolique, les pythagoriciens mêlaient à l'enseignement des mathématiques, de la musique et de la philosophie de la nature, un enseignement ésotérique. Ils encourageaient une vie communautaire, réservée aux initiés, mais porteuse d'un projet politique pour l'ensemble de la société. Celui qui souhaitait atteindre la vérité auprès du Maître Pythagore n'était pas seulement placé en face d'un discours théorique. Il ne devait pas uniquement s'efforcer d'adopter une droite méthode et de se libérer des préjugés. Son apprentissage exigeait aussi de signer la mort du profane qu'il était, de s'abandonner à une transformation intérieure, à changer son regard sur le monde. L'on pourrait dire avec Foucault que la vérité ainsi conçue « n'est donnée au sujet qu'à un prix qui met en jeu l'être même du sujet. Car tel qu'il est, il n'est pas capable de vérité²⁵. »

La transformation de soi au travers de la connaissance des premiers principes n'est évidemment pas étrangère à Platon. Il est plausible de penser qu'elle participe de cette définition de l'intuition intellectuelle qu'il donne dans la *République*. Ce qui est en jeu, en effet, c'est l'existence d'une faculté permettant de contempler la vérité au-delà de la science — cette dernière n'étant que le marche-pied d'une réalité supérieure. On trouve cette idée superbement expliquée dans cette tirade de Socrate :

Quel est donc de tous les dieux du ciel celui [...] dont la lumière permet à nos yeux de voir de la meilleure façon possible ? [...] C'est le soleil évidemment que tu me demandes de nommer. [...] Comme on a raison de penser que la lumière et la vue sont semblables au soleil, mais tort de croire qu'elles sont le soleil, de même, dans le monde intelligible, il est juste de penser que la science et la vérité sont l'une et l'autre semblables au bien, mais faux de croire que l'une ou l'autre soit le bien¹.

Chez Platon, les idées mathématiques, bien qu'abstraites, n'ont pas le même statut que les idées du Bien, du Beau ou du Vrai. Les premières se tiennent en effet dans le domaine du déductible, ce sont des principes ou des hypothèses. Les secondes relèvent de la contemplation, ce sont des vérités anhypothétiques ou inconditionnelles. Autrement dit, l'intuition contemplative est tenue pour supérieure à la logique.

L'histoire récente de la philosophie a, pour ainsi dire, inversé cet état de choses, pour faire des mathématiques, sinon le fondement de toute connaissance possible, tout au moins le modèle de scientificité. Cette attitude est favorablement capturée dans l'expression « La reine des sciences », qui échoit

25. FOUCAULT [?], *L'Herméneutique du Sujet*, p. 9.

aux mathématiques, laissant les philosophes-rois débattre de leur éviction. Il est cependant intéressant de se demander comment on en est venu à abandonner cette idée d'une intuition intellectuelle ?

9.5.2 La Critique kantienne

Au rang des auteurs ayant porté un coup fatal à l'intuition intellectuelle, on ne saurait négliger le rôle de Kant. Dans la critique de la raison pure, la typologie des facultés humaines, allant de la sensibilité à la raison, en passant par l'Entendement, épouse en quelque sorte l'analogie de la ligne de Platon, mais en faisant l'économie du quatrième terme. Ce n'est pas que la *dianoia* ait proprement disparu avec l'apparition du sujet transcendantal kantien, mais tout se passe comme si l'épaisseur anthropologique de l'être platonicien avait été mise à plat. Le concept de Raison, chez Kant, est la faculté de chercher, au sein de l'esprit, les principes inconditionnels sur lesquels reposent les catégories de l'Entendement. La Raison est donc l'analogue de la *dianoia* chez Platon, à cette différence près qu'elle ne s'opère pas par une quelconque contemplation de la vérité, mais par le truchement du raisonnement pur. Or ce raisonnement pur, nous dit Kant, en ce qu'il ne peut pas s'appuyer sur un critère de vérité qui lui serait extérieur, ne conduit qu'aux antinomies.

Précisons cette idée. Chez Platon, les Idées sont les objets de la pensée par excellence. Ces Idées existent indépendamment de l'esprit humain. Elles doivent donc déjà être données pour pouvoir être contemplées. Dans ces circonstances, il est difficile de soutenir la possibilité d'une connaissance strictement *a priori*, comme celle des mathématiques. C'est pourquoi la véritable méthode en philosophie spéculative, selon Kant, consiste à poser que ce sont les objets qui se règlent sur la connaissance humaine, et non l'inverse. Ce renversement de conception est une véritable révolution copernicienne : « Si l'intuition se réglait nécessairement sur la nature des objets, je ne vois pas comment on en pourrait savoir quelque chose *a priori* ; que si, au contraire, l'objet (comme objet des sens) se règle sur la nature de notre faculté intuitive, je puis très bien alors m'expliquer cette possibilité²⁶ » Tout ce qui fait l'objet de connaissance *a priori* se trouve toujours déjà en nous, préalablement à toute expérience sensible ou intellectuelle. Avec Kant, l'intuition intellectuelle (noétique) est remplacée par le jugement synthétique *a priori*.

Une conséquence majeure du tournant kantien, c'est qu'il n'est plus possible de connaître les choses autrement que comme phénomènes. Dans la mesure en effet où ce sont nos concepts qui se règlent sur les objets, et non l'inverse, nous ne connaissons jamais les objets que dans la modalité de notre perception. Dès lors, il est impossible d'étendre le savoir sur d'autres choses que des objets d'une expérience possible. Or il existe chez l'homme

26. KANT, *Critique de la raison pure*, trad. de BARNI, Préface de la seconde édition, p. 24.

une disposition naturelle, dans le processus de développement continu de sa connaissance, à poser des questions qui excèdent toute représentation intuitive. Ces questions sont celles touchant à Dieu, au monde et à l'âme. Le monde est-il fini ou infini ? A-t-il un commencement dans le temps ? A-t-il été créé par un Être Suprême ? L'âme est-elle une substance simple et incorruptible ? Ce sont ces interrogations qui ont nourri tous les systèmes de croyances, qu'ils soient métaphysiques ou cosmogoniques. Dans la mesure où de telles idées ne se réfèrent à aucune intuition pure ni à aucune expérience sensible, elles conduisent toujours à des raisonnements paradoxaux et, pour cette raison, doivent être délaissées. C'est l'objet de la *dialectique transcendantale* que de discuter précautionneusement chacune des conclusions logiques où mènent ces considérations, pour montrer leur situation aporétique inextricable.

À bien considérer les choses, le premier résultat de cette critique à laquelle s'attelle le philosophe de Königsberg est somme toute *négatif* : il s'agit de montrer que l'usage de la raison pure ne peut pas mener à une connaissance proprement scientifique. Kant proscriit donc la *noësis* de Platon, conçue comme la recherche de l'anhypothétique. Cette restriction cependant, loin de détrôner la dialectique du piédestal où la philosophie de Platon l'avait hissée, contribue tout au contraire à asseoir sa position. Certes il lui faut descendre quelques marches, certes il lui faut abandonner ses prétentions à embrasser d'un seul regard la totalité du savoir, mais l'exercice de la dialectique en tant que telle est sauvegardé : « il y a toujours eu dans le monde et il y aura toujours une métaphysique, mais toujours aussi on verra s'élever à côté d'elle une dialectique de la raison pure, car celle-ci lui est naturelle. La première et la plus importante affaire de la philosophie est donc d'enlever une fois pour toutes à cette dialectique toute pernicieuse influence en détruisant la source des erreurs²⁷ »

Il semble dès lors judicieux, plutôt que d'aller à rebours de la critique kantienne, de se demander si, dans notre monde frappé du sceau scientifique, il reste possible de justifier l'existence d'une faculté comparable à l'intuition intellectuelle ? Cette intuition devrait faire l'objet d'une définition mathématique rigoureuse, et reposer toute entière sur le jugement déductif, mais n'être en aucune manière susceptible d'être reproduite par une machine. Bien qu'une telle idée semble audacieuse, elle n'a, lorsqu'on prend la peine de l'envisager sérieusement, rien de proprement contradictoire. Nous manipulons au quotidien des objets qui échappent, par leur nature, au fonctionnement paradigmatique de l'informatique. À titre d'exemple, nous pouvons penser à la musique analogique, ou de manière générale, aux phénomènes physiques de l'ordre du continu, qui nécessitent d'être discrétisés (ou numérisés) pour être enregistrés et modélisés sur un ordinateur. Ne se pourrait-il pas que l'esprit humain, en ce qu'il a d'intuitif, soit capable d'appréhender

27. KANT, *op. cit.*, Préface de la seconde édition, p. 36

le continu ?

C'est cette idée, exposée ici de façon encore embryonnaire, que nous allons développer dans les chapitres suivants. Elle se présente, face au dilemme de l'intuition, comme une tierce voie. Plutôt que de défendre que l'intuition n'est pas définissable dans le langage, ou qu'elle ne l'est que dans une théorie qui soit en-deçà, ou au-delà, des mathématiques, nous affirmerons la chose suivante : *le concept d'intuition est explicitable au travers d'un langage qui, quoique mathématique, n'est pas implémentable par une machine.*

Cette posture nous conduira, tout d'abord, à définir le fonctionnement d'un ordinateur en exposant les principes de la machine de Turing et de la thèse de Church-Turing, ainsi que les principaux résultats limitatifs à la puissance de calcul de l'ordinateur. Une fois posées les définitions mathématiques de l'automatisation de la pensée, il conviendra de se demander quelles sont les failles de ces procédures ; ou autrement dit, de déterminer quels types d'opérations, pourtant mathématiques, ces machines ne parviennent pas à exécuter. Est-il possible que ces opérations soient précisément le propre de l'intelligence humaine ? Dans le cas contraire, peut-on dire, de source sûre, que l'intuition soit imitable par une machine ?

Lorsque l'on tente de définir pourquoi la machine manque d'intuition, on se confronte à un dilemme identifiée historiquement par Alan Turing :

- Soit on définit explicitement ce qu'est l'intuition au moyen de règles opératoires, mais dans ce cas, on ne comprend pas pourquoi une machine ne pourrait pas apprendre ces règles.
- Soit on ne définit pas l'intuition en des termes explicitement opérationnels, mais dans ce cas, on encourt le risque de ne pas savoir ce qu'est l'intuition au juste.

Dans ce chapitre, j'ai discuté deux approches pour contourner ce dilemme :

1. Au sein du *mystérianisme*, on suppose que la définition opérationnelle existe, mais que celle-ci est cognitivement close : il est donc impossible, pour nous autres humains, d'y avoir accès.
2. Au sein du *mysticisme*, il existe une définition explicite, mais celle-ci ne s'exprime pas dans le langage de la raison. C'est une définition, tout au plus, symbolique.

L'objet du prochain chapitre consistera à explorer une tierce voie : donner à l'intuition une définition explicitement opératoire, de telle sorte cependant que les machines ne puissent pas l'implémenter. En termes techniques, on dira que la fonction doit être *bien fondée* mathématiquement, mais *non calculable*.

Chapitre 10

La Machine de Chuck Norris

Chuck Norris already compted
at infinitum. Twice.

Anonymous

Nous avons vu que toute explication de ce qu'est l'intuition se heurtait à un dilemme de taille : soit cette explication est suffisamment claire et opérationnelle pour être décrite par une fonction mathématique, dans quel cas on ne voit pas d'objection de principe à ce qu'une machine de Turing parvienne à l'implémenter ; soit au contraire l'explication souffre d'ambiguïté, si bien qu'on n'est pas certain de savoir de quoi l'on parle exactement. J'ai montré que deux des stratégies classiquement envisagées par la philosophie pour contourner ce problème se heurtaient à des objections notables. Ni le mystérianisme ni l'hypothèse d'une intuition intellectuelle ne semblent apporter une réponse satisfaisante.

Dans ce chapitre, j'aimerais envisager une tierce voie : donner une définition à l'intuition humaine qui, bien que mathématiquement explicite, ne puisse pas être implémentée par un ordinateur. Cette réponse consiste donc à dire que l'esprit humain peut réaliser des opérations mathématiques qui échappent pourtant à la puissance de calcul des machines de Turing que nous avons vues jusqu'à présent (voir [5.3.1](#)).

Il est important, comme on l'a dit, que ma définition de l'intuition soit exprimable dans un formalisme mathématique. Or il semble difficile de mathématiser une notion aussi vague que celle d'intuition sans faire des hypothèses scientifiques sur le fonctionnement de l'esprit. Si je dis, par exemple, que l'intuition consiste en un ensemble d'informations qui, bien que traitées par le cerveau, ne le sont pas de façon conscientes, cela présuppose, d'une part, que l'esprit humain puisse être représenté comme une machine capable de traiter l'information et, d'autre part, que l'esprit humain s'exerce à travers le cerveau. Une telle vision s'accommode aisément avec le paradigme computationnaliste, mais elle pourrait être critiquée par les partisans de

l'énactivisme ou de l'esprit étendu, qui cherchent à expliquer la cognition en mobilisant, en plus du cerveau, le corps et sa relation à l'environnement.

Pour contourner cette difficulté, je ne chercherai pas à construire un modèle mathématique de l'intuition humaine, mais un modèle des fonctions mathématiques de l'intuition humaine. Quelle différence ai-je en vue exactement lorsque je parle de modèles et de fonctions ? Tout le monde accepte, je crois, l'idée que l'être humain puisse opérer des calculs sur les nombres. L'addition, la soustraction, la multiplication et la division sont des fonctions mathématiquement bien définies, qui peuvent être exécutées par un esprit humain, un abaque ou une calculatrice. Cela ne veut toutefois pas dire que l'esprit humain fonctionne de la même façon qu'un abaque ou une machine à calculer. D'un point de vue terminologique, on dira que l'addition est une fonction, tandis que l'esprit humain, l'abaque et l'ordinateur sont des modèles capables d'exécuter cette fonction.

Lorsque l'on dit que l'esprit humain se comporte comme un ordinateur, on ne veut souvent rien dire d'autre que ceci : les fonctions que l'esprit humain peut exécuter sont aussi satisfaites, dans une large mesure, par nos ordinateurs. Il existe toutefois des différences de fonctionnement notables : là où l'esprit humain est fait de cellules gliales, de myélines, de nœuds de Ranvier, l'ordinateur quant à lui est tressé de câbles et de plaques d'époxyde, de transistors, de condensateurs, de microprocesseurs. Là où l'ordinateur exécute des opérations séquentielles, binaires et rythmées par une horloge interne, le cerveau excite en parallèle, au moyen de signaux continus, un grand nombre de régions cérébrales interconnectées.

L'ensemble des fonctions qu'un ordinateur peut exécuter est précisément donné par le formalisme de la machine de Turing. Un ordinateur est en effet, par essence, un système logique capable de réaliser des calculs. La grande idée d'Alan Turing et d'Alonzo Church, appelée la thèse de Church-Turing, stipule que toute fonction calculable est exécutable par une machine de Turing (voir à cet effet l'Annexe 3). Autrement dit, tout ce que la machine de Turing peut faire, c'est du calcul. La question qui se pose est celle de savoir si l'esprit humain est capable de réaliser autre chose que du calcul. Ou autrement dit, s'il existe des fonctions mathématiques qui semblent échapper au formalisme de la machine de Turing, mais sans être pour autant hors de portée de l'intelligence humaine ?

Dans ce chapitre, je vais commencer par distinguer plus rigoureusement la différence existant entre un modèle de l'esprit humain et une fonction de l'esprit humain. Je vais également anticiper des objections qui seront discutées plus amplement dans la partie suivante de la thèse. Ensuite, je vais suggérer trois lignes stratégiques permettant de construire une fonction dépassant les capacités de calcul de la machine de Turing.

10.1 Fonctions et Modèles

Pour commencer, il semble crucial de bien distinguer ce que l'on entend par *fonction* et par *modèle*. La *fonction* est une relation qui associe, à toute valeur donnée, une réponse correspondante. Elle s'apparente à la Sybille des anciens Romains. Vous lui posez une question ; elle fournit la réponse. C'est bien d'ailleurs le sens donné à la définition mathématique. Les élèves apprennent dès l'école secondaire qu'une fonction est une transformation qui prend un certain objet du domaine de définition (par exemple un nombre naturel) et renvoie, pour chacun d'eux, un autre objet du domaine des images (par exemple un autre nombre naturel). Rien n'interdit cependant cette fonction de travailler, non pas avec des nombres, mais avec des fruits et des légumes. Il peut s'agir d'une transformation qui vous dit quel légume acheter compte tenu de la saison, par exemple :

<i>printemps</i>	$\rightarrow$	<i>avocat</i>
<i>été</i>	$\rightarrow$	<i>fraise</i>
<i>automne</i>	$\rightarrow$	<i>poire</i>
<i>hiver</i>	$\rightarrow$	<i>citron</i>

Le *modèle*, si on me permet de reprendre la métaphore mythologique, est une description qui cherche à expliquer ce qui se passe dans la tête de la Sybille lorsqu'elle donne une réponse à une certaine question. Le modèle essaie de comprendre d'où lui vient cette intuition, cet art visionnaire, qui lui permet de prévoir l'avenir. Il s'agit donc d'une théorie explicative élucidant, de façon rationnelle, le lien entre les entrées et les sorties d'une certaine fonction.

Pour récapituler ce qui vient d'être dit, la fonction peut être décrite par une boîte noire associant un input à un output, tandis que le modèle est un mécanisme, doté de règles de fonctionnement, capable d'exécuter une ou plusieurs fonctions. Cette distinction nous permet de comprendre que l'ordinateur et l'esprit humain possèdent certes une fonction en commun, qui est celle de calculer, mais sans pour autant que l'ordinateur soit un modèle valide de l'esprit humain. Ce que je veux dire par là, c'est que si vous donnez une opération arithmétique à un homme rompu aux calculs mentaux, il va produire — on l'espère du moins — une réponse identique à celle de la machine. La même équation conduit toujours à la même réponse. Il est toutefois douteux de penser que l'esprit humain et l'ordinateur présentent les mêmes mécanismes de fonctionnement internes. Autrement dit, la façon dont la réponse est produite ne suit pas forcément les mêmes règles.

Quelles sont exactement les différences de fonctionnement existant entre le modèle de l'ordinateur, qui n'est autre que la machine de Turing (voir section 5.3.1), et le modèle de l'esprit humain ? Remarquons que l'être humain, contrairement à l'ordinateur, peut commettre régulièrement des erreurs de

raisonnement — surtout lorsque le calcul excède sa mémoire de travail. Par ailleurs, il n'est pas sûr que « calculer » soit la fonction la plus essentielle de l'esprit humain, alors que c'est bien la fonction principale de l'ordinateur. Il existe en effet des cultures où le concept de nombre est inexistant. Il existe également des individus présentant une dyscalculie. Que vous soyez dyscalculique ou que vous veniez d'une tribu primitive, cela ne vous empêche pas d'être doté d'un esprit humain capable de raisonnement, de cognition sociale, d'adaptation, etc. In fine, il se pourrait que le calcul, entendu comme la capacité à produire de raisonnement strictement logico-déductif, ne représente qu'une facette, somme toute secondaire, des aptitudes humaines.

Dans ce qui va suivre, j'aimerais me demander en quoi précisément l'esprit humain et la machine, envisagés comme des modèles de calcul, peuvent différer.

10.1.1 Effet SNARC : structure intuitive de l'arithmétique

Comment fonctionne exactement le calcul humain ? Quels ingrédients utilise-t-il qui le différencie de la machine ? Si vous interrogez votre entourage, vous découvrirez peut-être que certaines personnes font des calculs mentaux sans pouvoir dire d'où leur vient la réponse. Tout se passe pour eux comme si le résultat s'imposait de façon inconsciente. D'autres individus, au contraire, visualisent des chiffres dans leur tête et les manipulent mentalement. D'autres encore présentent la particularité de se représenter les nombres dans l'espace visuel ou péripersonnel. Ces nombres possèdent pour eux une position dans l'espace ; ils flottent dans les airs, il est possible de les pointer du doigt. D'autres, pour finir, voient des couleurs. Ils associent chaque nombre à une tonalité, voire même à un son ou à un personnage. Les opérations arithmétiques sont pour eux des fresques colorées, des morceaux de musique, des expériences sensorielles.

En dépit de la diversité des expériences, est-il possible de trouver un modèle commun pour expliquer la façon dont l'esprit humain procède à des opérations arithmétiques ? Des études récentes suggèrent la présence d'une structure visuo-spatiale implicite, consistant à disposer les nombres sur une ligne allant de gauche vers la droite (du moins dans les cultures occidentales^[1]). Pour obtenir ce résultat, un dispositif expérimental est mis au point. Au sein de ce dernier, chaque sujet est confronté à deux nombres compris entre 0 et 9 ; l'un à gauche de l'écran, l'autre à droite. Il doit dire lequel des deux nombres est plus grand. De façon intéressante, il a été remarqué que le temps de réponse était plus court lorsque le nombre à droite était plus grand que celui de gauche. Ce biais est ce qu'on appelle l'effet SNARC.

L'effet SNARC suggère qu'il y a une tendance automatique, et pour ainsi dire naturelle, à associer les nombres à une position dans l'espace.

1. DEHAENE ET AL., *The mental representation of parity and number magnitude*, J. Exp. Psychol., 122, 371–396, 1993

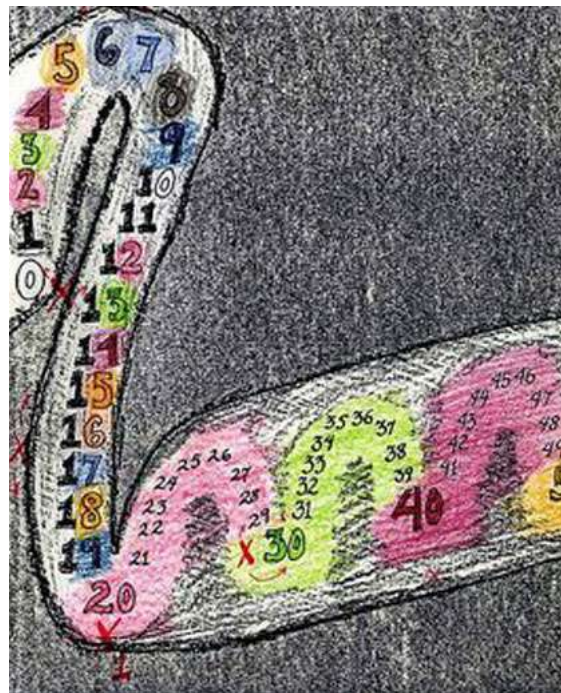


FIGURE 10.1 – Voici la représentation donnée par un enfant présentant une synesthésie numérique spatiale.

Cette visualisation spatiale est consciente chez certaines personnes ; elle est inconsciente chez les autres. Les synesthètes, par exemple, sont capables de dessiner la position des nombres dans l'espace. Il leur est même impossible de ne pas penser aux nombres indépendamment de cette structure spatiale. La figure 10.1 donne une illustration de la façon dont un synesthète est susceptible de se représenter l'échelle des nombres naturels. On remarque que ces nombres sont orientés de la gauche vers la droite, conformément aux résultats supportés par l'étude de l'effet SNARC. Cette orientation est toutefois loin d'être linéaire : on observe des convolutions qui semblent défier l'imagination².

Il est permis de croire, à titre d'hypothèse, que cette visualisation spatiale des nombres rend l'être humain capable de raisonner sur des grandeurs infinies. Lorsque l'on s'exprime à propos de la suite des puissances de deux, par exemple, l'on se représente une grandeur spatiale qui est toujours deux fois plus grande que la grandeur précédente, jusqu'à aboutir asymptotiquement à la notion de droite infinie. La possibilité de cette droite est

2. Dans cet exemple, chaque nombre, en plus d'avoir une position dans l'espace, est aussi associé à une couleur. L'association d'un nombre et d'une couleur est appelée synesthésie graphème-couleur, tandis que la position des nombres dans l'espace est la synesthésie numérique spatiale. L'un et l'autre ne vont pas toujours de pair.

visuellement claire, bien qu'il soit impossible de l'appréhender dans toute son extension. On pourrait dire que la représentation spatiale des nombres est une structure perceptive *a priori*, dans laquelle notre compréhension intuitive des opérations arithmétiques prend place. Cette structure présente un ordre bien fondé. Il y a des nombres qui sont « avant » et d'autres qui sont « après ». Il y a également un « début », un point initial, mais il n'y a pas de limite au processus, qui s'étire sans fin jusqu'à l'infini³.

Un ordinateur ne semble pas disposer de cette structure intuitive. Il n'associe pas les nombres à une grandeur spatiale de façon automatique. La preuve en est que l'ordinateur n'est pas victime de l'effet SNARC. Il n'y a pas de raison de croire que si on lui présente des nombres sur une feuille de papier, la position de ces nombres les uns par rapport aux autres influence sa vitesse d'exécution de la même façon que chez l'être humain. Ce que je veux dire par là, c'est qu'il est possible d'écrire un programme informatique qui s'affranchisse de l'effet SNARC, tandis qu'il est impossible, pour un être humain, de ne pas être victime de ce phénomène.

L'absence de visualisation spatiale intrinsèque des nombres pourrait expliquer pourquoi la machine est incapable de donner une signification à l'infini. L'ordinateur peut certes manipuler des nombres tendant asymptotiquement vers l'infini, mais peut-il se représenter cette ligne infinie vers laquelle tend la série de ces nombres ? Ce qui est spécifique à la visualisation humaine d'une droite, c'est que l'idée de l'infini en acte et de l'infini en puissance se rencontrent, pour ainsi dire, automatiquement. Ce que je veux dire par là, c'est que l'esprit humain se rend capable de concevoir qu'une suite de nombres croissants converge vers quelque chose qu'il nomme l'infini. Il se représente un concept, celui d'une droite infinie, qui se situe au-delà de toute représentation possible. Il peut donner un sens visuel à un concept pourtant abstrait.

De ces quelques remarques, on conclut deux choses : premièrement, que la machine à calculer n'est pas forcément le modèle de l'esprit humain, pas plus que la machine de Turing ne serait le modèle du cerveau. Ce que l'être humain et la machine ont en commun, ce n'est pas la façon de traiter l'information, mais le fait de pouvoir exercer un certain nombre de fonctions — dont le fait de calculer.

De plus, et il s'agit de la deuxième remarque, il se peut que le fonctionnement de l'esprit humain lui permette d'implémenter des fonctions qui n'existent pas dans une machine à calculer ni dans une machine de Turing. On a épinglé, notamment, le fait de visualiser l'infini, ou de comprendre certaines propriétés relatives à des suites tendant vers l'infini. Il se pourrait

3. Cette idée qu'il existe une structure intuitive de la perception a déjà été discutée dans le chapitre 9. J'y ai défendu l'idée que, contrairement à un programme informatique, l'être humain est sujet à des illusions d'optiques. L'intuition peut être conçue comme des associations automatiques influençant le jugement, des structures liant des concepts les uns aux autres, et ce indépendamment de notre volonté.

donc que, en explorant plus avant cet aspect, l'on découvre qu'il existe une fonction, dans le chef de l'esprit humain, qui requiert un autre modèle que la machine de Turing classique. C'est pourquoi, dans la section suivante, je vais m'intéresser à ce à quoi ressemblerait une machine de Turing si on la dotait de la capacité de se représenter l'infini ?

10.1.2 Vers un modèle de calcul transfini

L'histoire de la philosophie de l'intelligence artificielle est jonchée de déclarations, à propos de l'intuition, qui se sont avérées inexactes. Hubert Dreyfus disait que la machine ne pourrait pas vaincre l'être humain aux échecs, sous prétexte qu'elle ne disposait pas de la capacité d'apprentissage intuitive du jeu, permettant de reconnaître des schémas de façon instantanée⁴. Les représentants de l'école de Palo Alto, pour leur part, affirmaient qu'une machine ne pouvait pas s'orienter dans une ville, à cause que cela demandait une compréhension bien trop fine du contexte. Ces exemples, aujourd'hui obsolètes, démontrent que les méthodes informatiques ont régulièrement démenti les réserves fondées sur l'intuition.

Est-il néanmoins possible de formuler une objection de principe ? Est-il permis de construire un argument qui soit indépendant des conditions particulières de l'informatique de notre temps ? Cela supposerait de montrer, par A plus B , que l'intuition dont il est question (apprentissage, sensibilité au contexte...) est inimitable par une machine de Turing. Il y a plusieurs façon de procéder :

- **Stratégie fondée sur l'indécidabilité** : Une première approche consiste à définir l'intuition de façon à ce que la fonction satisfaite par celle-ci soit mathématiquement réductible à un problème indécidable. Au nombre de ces problèmes se trouve notamment la question de savoir si une équation diophantienne possède une solution. Une équation diophantienne est une équation polynomiale n'admettant que des solutions entières (des nombres sans virgules). Le théorème de Youri Matiassevitch a en effet montré qu'il était impossible de construire un algorithme sur une machine de Turing qui répondrait à cette question pour n'importe quelle équation diophantienne. Il est seulement possible de produire des algorithmes fonctionnant au cas par cas. Si donc la définition de l'intuition est « la capacité à déterminer, dans le cas le plus général, si une équation diophantienne admet une solution », on tiendrait à coup sûr quelque chose qu'une machine ne pourrait pas faire. Il ne resterait plus qu'à argumenter que l'être humain, lui, peut y arriver ! Ce genre de stratégies a été popularisée par l'argument de Lucas-Penrose, qui repose sur une astuce

4. DREYFUS, *What computer still can't do ?*

quelque peu similaire, si ce n'est qu'il utilise les théorèmes de Gödel et non pas les machines de Turing. Une des principales objections auxquelles il faut faire face, lorsque l'on adopte ce genre de stratégie, c'est la faillibilité de l'intuition. L'intuition est en effet rarement présentée comme une aptitude adéquate à tous les cas de figure. Elle est sujette à l'erreur, mais également à l'absence de réponse. Elle ne correspond donc que difficilement à « la capacité à déterminer *à coup sûr*, dans le cas le plus général, si une équation diophantienne accepte une solution ».

- **Stratégie fondée sur le hasard** : Si on prend au sérieux le caractère aléatoire, fulgurant et capricieux de l'intuition, on pourrait être tenté de la modéliser à l'aide du hasard ? Les algorithmes aléatoires ne font pas partie de l'équipement d'une machine de Turing classique, mais recourt soit d'en appeler à des machines de Turing non déterministes, soit de recourir à des ordinateurs quantiques. Certains algorithmes quantiques montrent en effet que l'usage du hasard permet d'augmenter la résolution des problèmes appartenant à la classe des problèmes NP-complets, réputés parmi les plus difficiles en mathématiques (voir section 6.6). La résolution d'un tel problème par force brute, en vérifiant une à une chacune des solutions possibles, demande un temps exponentiel dans le pire des cas. Un algorithme quantique permet de réduire ce temps par une racine carrée. Cela reste exponentiel, mais tout en apportant un gain de rapidité. L'intuition humaine agirait-elle différemment ? Le problème central avec cette stratégie argumentative, c'est qu'on pourra arguer que l'intuition n'est qu'une affaire de paradigme informatique. Dès qu'on aura des ordinateurs quantiques, il n'y aura plus lieu de penser que l'intuition soit une faculté humaine.
- **Stratégie fondée sur l'infini** : C'est pourquoi la ligne argumentative la plus générale qui puisse être envisagée consistera à dire que l'intuition, et particulièrement l'intuition de l'infini dont il était question plus haut, nécessite de calculer l'infini. Qu'est-ce que je veux dire par là ? Qu'il soit possible d'atteindre la limite d'un processus infini, tel par exemple que l'énumération 1, 2, 3, ... Cette aptitude est au-delà d'une machine de Turing et, vraisemblablement, de toute machine concevable en pratique. Il est néanmoins permis de construire une extension des machines de Turing : les machines de Turing transfinie.

La tâche qui se présente à moi est plurielle. D'une part,(1) il s'agira de montrer que les fonctions que je viens de définir (l'incalculabilité, le hasard, l'infini) sont celles dont se rendent justement capables l'esprit humaine.

D'autre part, (2) il s'agira également de construire un modèle de ces fonctions. Ce modèle sera conçu comme une extension transfinie des machines de Turing classique. C'est-à-dire, une machine susceptible, tout comme Chuck Norris, de compter l'infini... plusieurs fois d'affilée.

Ce programme dont j'annonce la couleur est plus ambitieux qu'il ne peut paraître à prime abord. Construire un modèle qui réunit en lui les deux intuitions (celle du Tout et de l'Infini) et les trois fonctions (indécidabilité, hasard et infini) revient à démontrer l'existence d'une équivalence conceptuelle entre ces multiples notions. Plus précisément, cela suppose l'existence d'une unité mathématique X liant les problèmes indécidables, les problèmes NP-complets et les séries infinies, puis à construire explicitement un modèle qui supporte X . Loin de se cantonner uniquement à la philosophie, ce projet est donc susceptible de nous apprendre également quelque chose en mathématique. Je ne prétends pas démontrer des résultats nouveaux, mais offrir, sur des problèmes connus, un regard alternatif. Cet élément X énigmatique, c'est précisément la machine de Turing transfinie.

Avant de me lancer dans la définition formelle de la machine de Turing transfinie, et d'en étudier les principales propriétés, j'aimerais commencer par justifier que les trois fonctions que j'ai énumérées ci-dessus capturent effectivement les intuitions de l'Infini et l'intuition du Tout. Plus précisément, je voudrais justifier que l'intuition de l'Infini (l'intuition transcendante) est liée à deux fonctions : la complétion de séries infinies et la résolution de problèmes indécidables ; tandis que l'intuition du Tout (l'intuition holistique) est liée à la résolution des problèmes NP-complets (voir tableau ci-dessous).

Définitions	Fonctions
Intuition transcendante	Compléter des séries infinies Solutionner des problèmes indécidables
Intuition holistique	Solutionner des problèmes NP-complets

10.1.3 L'intuition de l'Infini face aux problèmes indécidables

Il est aisé de concevoir que l'intuition de l'infini permette de se représenter la convergence de séries infinies. C'est du moins ce en faveur de quoi j'ai plaidé au moment de parler de la synesthésie et de l'effet SNARC. En revanche, il n'est pas particulièrement évident de comprendre pourquoi la résolution de problèmes indécidables ou de problèmes NP-complets serait lié, d'une manière ou d'une autre, à cette intuition de l'infini.

Je vais donner ici un éclaircissement de ce point, qui recevra ultérieurement de plus amples justifications. Il s'agira de dire que la résolution des problèmes indécidables est liée à l'intuition de l'Infini, tandis que la résolution de problèmes NP-complet est liée à l'intuition du Tout.

J'aimerais tout d'abord suggérer qu'une machine capable de compter

l'infini est sans doute aussi en mesure de trouver la solution à des problèmes indécidables. Afin de comprendre ce lien entre infini et indécidabilité, il peut être judicieux de rappeler certains résultats mathématiques ? Lorsque les premiers problèmes indécidables ont été découverts par Gödel et par Turing, les logiciens se sont penchés sur la question de savoir comment les contourner. Cela a donné naissance à une clarification rigoureuse des hypothèses sous-jacentes au théorèmes de Gödel. Ces hypothèses, dans les ouvrages vulgarisés de logique, sont trop souvent ignorées. C'est pourquoi j'ai pris la peine de les rappeler dans l'annexe 2).

L'une de ces hypothèses nous intéresse au plus haut point. Un mathématicien du nom de Gentzen est parvenu à la conclusion que si l'on disposait d'un système de preuve transfini, les indécidabilités seraient résolues. Dit autrement, l'indécidabilité vient du fait qu'une preuve mathématique est de taille finie. On commence avec des hypothèses au début du raisonnement, on procède à une succession d'étapes, avant d'atteindre un résultat final. Or il se trouve que si on permettait à la preuve de se poursuivre jusqu'à l'infini, il serait possible d'apporter une réponse aux problèmes indécidables. Il serait, en particulier, possible de prouver la cohérence de l'arithmétique.

Il ne faut donc pas être un expert en logique mathématique pour comprendre qu'une machine de Turing capable de compter l'infini serait sans doute en mesure de résoudre les problèmes indécidables. Je prouverai ce résultat en temps et en heure. Pour l'instant, j'aimerais encore toucher un mot sur le rapport entre l'intuition du Tout et les problèmes NP-complets.

10.1.4 L'intuition du Tout face aux problèmes NP-complets

La résolution de problèmes NP-complets, contrairement aux problèmes indécidables, ne semble pas requérir l'infini. Un exemple notoire de problème NP-complet est la résolution du jeu informatique Démineur, qui consiste à cliquer sur des cases en évitant les bombes. On comprend bien que ce jeu présente une certaine complexité, mais on voit difficilement en quoi une machine capable de compter jusqu'à l'infini serait plus efficace à ce jeu. On souhaite en effet résoudre ce jeu en un temps fini, si bien que notre capacité de compter l'infini ne nous avancerait pas tellement dans nos objectifs.

C'est pourquoi je défendrai que les problèmes NP-complets sont liés, non pas à l'intuition de l'Infini, mais à l'intuition du Tout. Il y a en effet quelque chose, dans les problèmes du type du Démineur, qui exige de construire une vue d'ensemble. La complexité calculatoire s'explique en raison du fait que l'information détenue à propos d'une case spécifique du plateau de jeu exerce une influence sur des cases pourtant fort éloignées. Tout se passe donc comme s'il fallait avoir à l'esprit l'ensemble des configurations possibles du jeu pour produire une réponse optimale. Or cette vue d'ensemble excède très vite les capacités de calcul d'un être humain. C'est pourquoi ce dernier va se fier à son expérience. Il va tenter de restreindre le problème en se concen-

trant sur un petit nombre de cases, mais toutes les fois où il lui manque de l'information, il va se référer à la reconnaissance de patterns. Un joueur de Démineur de haut niveau se fonde sur la forme des motifs pour guider ses choix.

Pour bien comprendre la structure générale d'un problème NP-complet, et la raison pour laquelle elle est liée à l'intuition du Tout, je démontrerai que le problème NP-complet peut recevoir une illustration géométrique visuelle. Plus précisément, j'affirmerai qu'un problème NP-complet consiste à trouver, à partir d'un certain graphique, les points où une fonction s'annule (voir illustration [10.2](#)). Pour l'être humain, il existe un effet de saillance qui fait que notre regard est spontanément attiré vers ces points. Il nous est donc permis de le trouver sans devoir balayer une à une chacune des abscisses. Or jusqu'à preuve du contraire, une machine de Turing ne semble pas en mesure de déterminer rapidement les points où la fonction s'annule : la vitesse des algorithmes, dans le cas le plus général, n'est pas significativement plus rapide que s'ils balayaient les abscisses une à une.

Pour reprendre l'exemple du Démineur, même nos meilleurs algorithmes connus risquent d'énumérer successivement toutes les possibilités : ils vont envisager le cas où une bombe se trouve sur la case A et la case B, le cas où une bombe se trouve sur la case A mais pas la case B, le cas où elle se trouve sur la case B mais pas la case A, et ainsi de suite. Le joueur humain de haut niveau ne peut pas se permettre d'étudier tous ces cas un par un. Il se référera à son expérience et à son instinct pour orienter son attention immédiatement vers les scénarios les plus intéressants. C'est-à-dire non seulement les scénarios les plus probables, mais aussi ceux qui vont lui permettre de résoudre plus rapidement le problème.

Cette capacité de saillance dont je parle n'est pas sans rappeler le fonctionnement de la perception humaine que j'ai évoqué dans l'analyse des illusions temporelles (chapitre [3](#)). La conscience est toujours conscience de quelque chose qui change et, pour cette raison, elle n'envisage jamais les points d'un graphique de façon isolée, purement séquentielle, mais elle les voit comme des ensembles partiellement distincts, partiellement comparables les uns aux autres. On pourrait dire que, sous son regard, les points de l'espace sont « intriqués ». Il ne voit pas chaque scénario de façon isolée, mais il associe automatiquement un scénario à un second. L'information se présente à lui sous une certaine structure préalable, que l'expérience lui permet d'affiner avec le temps.

10.1.5 Réponses à une objection préliminaire

Avant de terminer cette section, j'aimerais encore prendre la peine d'anticiper une objection cruciale. J'ai dit que seule une machine transfinie pouvait implémenter les fonctions de l'intuition. Or il peut sembler délicat de dire que l'intuition soit une faculté à appréhender l'infini alors que le cerveau

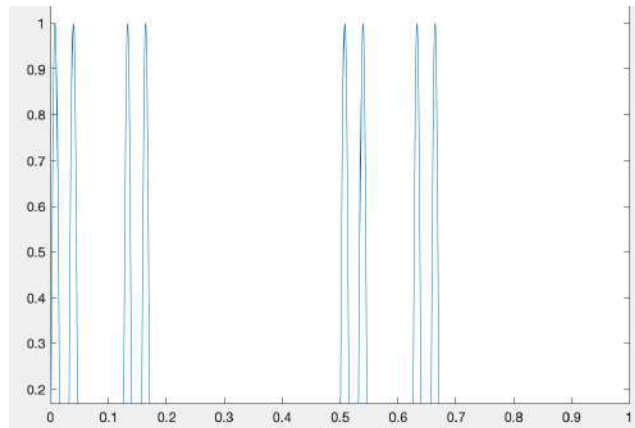


FIGURE 10.2 – Lorsqu'on lui donne une représentation graphique, un problème NP-complet consiste à identifier les points où une fonction s'annule. Dans l'exemple ci-dessus, il existe de nombreuses solutions, mais imaginez-vous un schéma plus compliqué n'admettant qu'une seule solution ? Pour peu que la taille du graphique soit d'une résolution suffisante pour l'œil humain, je postule que ce dernier est spontanément attiré vers les "creux". Il peut donc identifier la présence d'une solution en un temps "rapide", qui ne lui demande pas de balayer méthodiquement les abscisses une à une.

humain, lui, est fermement circonscrit dans la boîte crânienne. Comment un organe disposant d'un nombre fini de neurones, d'un volume délimité et d'une vitesse de traitement bornée peut-il se saisir de l'infini ?

Cette objection d'ordre méthodologique peut recevoir plusieurs réponses. Tout d'abord, il est important de noter que, bien que le cerveau soit fini, ce n'est pas forcément le cas de l'esprit lui-même ; réduire l'esprit humain aux capacités de calcul des neurones représente une pétition de principe, puisqu'il est alors supposé que l'esprit est quelque chose de calculable. Or c'est bien cette question qui est en jeu. Je montrerai dans la dernière partie de cette thèse que l'idée d'après laquelle l'esprit est dans le cerveau, aussi convaincante qu'elle puisse paraître, est on ne peut plus naïve. Je m'appuierai, pour conduire ma critique, sur une lecture philosophique des neurosciences.

Cela étant dit, j'accepte volontiers l'idée que l'esprit humain soit tout aussi fini que le cerveau. Il ne peut en effet pas se représenter l'infini en acte, c'est-à-dire l'infini dans sa totalité. Tout au plus peut-il s'en faire une idée asymptotique, en imaginant des étendues spatiales, des échelles de nombres, des échelles de temps infiniment croissantes. Il est donc capable de se représenter l'infini en puissance. Lorsque l'esprit est confronté à des séries de nombres, telles que 1, 1, 2, 3, 5, 8, 13..., il peut (1) tenter d'en prédire les régularités (ne serait-ce pas la suite de Fibonacci ?), visualiser que la suite diverge vers l'infini et se représenter la fin de cette énumération à l'aide du symbole de l'infini ; il peut même hiérarchiser les différentes formes d'infinis

comme dans la théorie de Cantor, et imaginer des infinis de taille toujours plus grands.

Pour finir mon plaidoyer en faveur des machines de Turing transfinie, remarquons qu'il existe de nombreux phénomènes, délimités dans le temps et dans l'espace, qui sont modélisés en recourant à des notions infinies. La géométrie euclidienne, par exemple, suppose que l'espace soit infini, alors même que notre Univers est de taille bien circonscrite. Les abstractions que sont les droites infinies (infiniment grand) et les points n'occupant aucun volume (infiniment petit) sont certainement hors de portée de notre expérience quotidienne de l'espace, mais ils n'en demeurent pas moins pertinents pour décrire cette expérience^[5]. Un autre exemple est celui de la thermodynamique, qui suppose l'existence d'un espace ouvert. Un tel espace n'est théoriquement possible que si l'univers était infini. Il y a également des espaces infinis dans la formalisation de l'espace de Hilbert, utilisé pour décrire les équations de la physique quantique. Pour finir, n'oublions pas que la machine de Turing classique est elle-même déjà une idéalisation. Sa capacité de mémoire et de traitement est supposée infiniment extensible^[6]. Cela ne correspond pas forcément à l'expérience de l'esprit humain, ni même aux ordinateurs concrets que nous manipulons, qui possèdent des ressources limitées.

10.2 La Machine de Turing transfinie

Dans la section précédente, je me suis contenté d'établir un cahier des charges. J'ai spécifié, point par point, certaines des propriétés dont une machine devrait se montrer capable pour satisfaire les fonctions de l'intuition. Le premier critère, celui qui a retenu ma préférence immédiate, consiste à construire une machine capable de déterminer la convergence d'une série infinie, même en l'absence de lois connues. Une telle machine devrait pouvoir dire que les suites $1 + \frac{1}{2} + \frac{1}{4} + \dots$ converge vers 2. Elle devrait en outre être capable d'énumérer une séquence de symboles alphabétiques du type

5. On pourra objecter que la géométrie euclidienne ne vise pas à modéliser l'Univers et qu'elle a été largement supplantée, dans la physique contemporaine, par des géométries non euclidiennes. D'après le modèle d'Einstein-Lemaître, par exemple, l'Univers possède un rayon de courbure variable, si bien qu'il oscille entre un espace elliptique et un espace hyperbolique ; ne prenant la forme d'un espace euclidien qu'à l'instant t , totalement éphémère, où le rayon de courbure devient nul. Cette objection ne met pas en danger le point qui est défendu, car l'espace hyperbolique lui-même suppose l'infini. Du reste, notre expérience de l'espace, sur Terre, est adéquatement modélisée par la géométrie euclidienne (en tout cas pour des petites distances) bien que la planète entière soit mieux représentée par la géométrie sphérique (qui est finie).

6. La machine de Turing est en effet dotée d'un ruban qui peut être prolongé aussi souvent que nécessaire. Ce ruban est utilisé pour consigner les symboles. Il sert donc d'espace de stockage, mais aussi d'espace de traitement. C'est, pour l'exprimer à l'aide d'une métaphore, la feuille de papier qui sert de support à ses calculs.

abcaaabcaa sorte à voir si cette séquence se « termine » par une suite infinie de *a* ou si elle boucle sans fin sur le motif *abcaa*.

La seconde stratégie, quant à elle, s'inscrit dans les brisées de l'argument de Lucas-Penrose. Elle consiste à construire une machine capable de dépasser certains résultats d'indécidabilité. Il peut s'agir des théorèmes d'incomplétude de Gödel, mais également du problème de l'arrêt, du théorème de Rice, etc.

La troisième stratégie, pour finir, consiste à doter la machine d'une certaine dose de hasard et d'espérer que cela puisse expliquer des différences de fonctionnement concernant, par exemple, la résolution de problèmes complexes du type P-NP. Il est en effet souvent admis que la capacité à trouver, par « chance », un bon candidat de solution à ce type de problème relève typiquement d'une forme d'intuition — là où la machine ne procède que par une suite de calculs déterministes.

Dans les pages qui vont suivre, j'aimerais proposer une révision de la machine de Turing permettant de satisfaire, au moins partiellement, chacun de ces trois critères. Bien que le modèle que je vais construire soit pensé premièrement pour « calculer l'infini » (le premier critère), je vais montrer qu'il possède des propriétés très utiles concernant les théorèmes d'indécidabilité (le deuxième critère) et la résolution des problèmes NP-complets (le troisième critère). Pour le dire autrement, j'aimerais bâtir un formalisme susceptible de capturer la notion d'intuition sous ses trois acceptions fonctionnelles.

Ma révision consistera à donner une extension à la définition classique de la machine de Turing, aussi bien en ce qui concerne la dimension spatiale que temporelle. Concernant la dimension spatiale, tout d'abord, je vais étendre la bande de lecture de sorte à ce qu'elle puisse contenir un nombre effectivement infini de symboles. Concernant la dimension temporelle, j'aimerais que la machine puisse réaliser un nombre effectivement infini d'opérations. Par « effectivement infini », j'entends que la machine peut compter l'infini et écrire 1, 2, 3, ... jusqu'à arriver au symbole ∞ . Elle peut ensuite poursuivre sa lecture pour écrire une nouvelle séquence de nombres 1, 2, 3, ..., et ce de façon récursive.

Cette approche que je me propose d'entreprendre soulève un certain nombre de difficultés. Il convient tout d'abord de se demander ce que cela signifie que de compter jusqu'à l'infini. Il est en effet facile de constater qu'il existe des infinis plus grands que les autres. Énumérer la suite des entiers naturels 1, 2, 3, ... semble plus aisé que de compter l'ensemble des nombres réels. Il y a en effet une infinité de nombre réels entre les nombre 1 et 2, mais également entre les nombres 1, 5 et 2, ou entre les nombres 1, 9 et 2. En fait, il est établi que, si vous prenez deux réels quelconques, vous aurez toujours une infinité de réels qui les sépare.

Cette idée que l'ensemble des réels $\mathbb{R}$ forme un ensemble plus grand que l'ensemble des nombres naturels $\mathbb{N}$ a été formalisée par Georges Cantor à la

fin du $XIX^{\text{ème}}$ siècle (voir l'annexe 4). Sa grande idée peut être résumée par une expérience de pensée. Supposez que vous ayez deux grands sacs contenant, chacun, un ensemble d'objets. Supposons en outre que vous vous mettiez marotte en tête et que vous demandiez : quel sac contient le plus d'objets ? Pour ce faire, vous prenez de votre main gauche un article dans le premier sac ; de votre main droite, vous en prenez-un dans le second sac. Il semble admissible de penser que, si les sacs ont la même taille, il n'y aura aucun moment où l'une de vos deux mains sera vide. Pour le dire autrement, vous pourrez toujours mettre en correspondance un objet du premier sac (main gauche) avec un objet du second sac (main droite).

En mathématique, on appelle cette opération, qui consiste à appairer les objets de deux ensembles, sous le nom d'une bijection. Cette notion semble efficace pour dire si un sac contient plus d'objets qu'un autre. Un sac contenant toutes les lettres de l'alphabet, par exemple, pourra être mis en bijection avec un sac contenant les nombres allant de 1 à 26, mais pas avec un sac contenant tous les mots du dictionnaire. L'idée de Cantor est d'étendre cette définition aux ensembles de tailles infinies. Sa question devient donc : est-il possible de mettre en bijection le sac contenant tous les entiers naturels ($\mathbb{N}$) avec le sac contenant tous les nombres réels ($\mathbb{R}$). Il se trouve que la réponse est négative : il existe infiniment plus de nombres réels que de nombres entiers. Dans le jargon mathématique, on dira que $\mathbb{N}$ est de cardinalité $\aleph_0$ tandis que $\mathbb{R}$ est de cardinalité $\aleph_0^2$ (voir annexe 5).

Une machine de Turing transfinie, notée T_{ω_1} , ou plus simplement T_1 , est une généralisation de la machine de Turing classique. Elle possède tout comme elle une tête de lecture, une bande infinie, un ensemble d'états et une fonction de transition, mais s'en distingue à deux égards. Tout d'abord, le ruban est de cardinalité infinie, et non pas seulement de taille non bornée. Plus précisément, nous imposerons que son ordinalité soit inférieure à ω_1 , mais en-dessous de cette borne, il pourra être rallongé aussi souvent que nécessaire. Par ailleurs, et c'est là le point le plus important, T_1 est capable de réaliser un nombre α d'opérations successives, pour peu que ce nombre α reste, lui aussi, strictement inférieur à l'ordinal ω_1 . Là où une machine de Turing classique tournerait indéfiniment, T_1 est capable d'atteindre la limite du processus récursif. Cela lui permet, notamment, d'écrire un à un chacun des entiers naturels, ou de s'arrêter après avoir atteint le « bout » de la bande (dans un sens qui va être précisé). Sous l'hypothèse du continu, T_1 peut également écrire un nombre irrationnel sur le ruban. En revanche, il lui est impossible d'énumérer l'ensemble des réels, ou même l'ensemble des points sur un segment de droite, car ces ensembles sont de cardinalité $\aleph_0^2$.

Contrairement aux machines dotées d'un "Oracle" (voir section 5.3.4), T_1 ne peut pas apporter, en une seule unité de temps, la réponse à des questions qui requerraient d'effectuer une infinité d'itérations. C'est plutôt nous autres observateurs qui sommes capables, lorsque nous envisageons une machine transfinie, de compter les ω_1 opérations nécessaires pour que

s'achève une récursion. Autrement dit, la définition d'une machine de Turing transfinie ne modifie pas tant les propriétés de la machine de Turing classique en elle-même, que notre façon de compter le temps, potentiellement infini, pendant lequel la machine s'exécute. Nous reviendrons plus en détails sur cette comparaison avec les machines de Turing munies d'un oracle, mais après avoir formalisé notre définition de machine de Turing transfinie.

L'extension de la machine de Turing, telle que je développe dans les pages qui vont suivre, s'apparente très étroitement aux modèles de l'Infinite Time Turing Machine (ITTM) et des Machines de Zénon. À la fin de ce chapitre, je vais exposer les différences formelles existant entre ma définition et celles existant dans la littérature.

10.2.1 Définition formelle

Considérons un ruban constitué d'un nombre infini dénombrable de cases indicées. Sur la première d'entre elles se trouve le symbole ineffaçable $\triangleright$. C'est à la fois l'extrémité gauche de la bande, mais aussi l'emplacement auquel commence la tête de lecture. Chaque fois que $\mathbf{T}_1$ lit ce symbole, la fonction de transition lui donne l'instruction $\rightarrow$, lui ordonnant de se rendre une case sur la droite, avec une réaffectation éventuelle (mais non obligatoire) de son état actuel. Dans certains cas, la machine peut boucler indéfiniment entre le symbole $\triangleright$ et le deuxième emplacement de la bande. Pour une machine de Turing transfinie, ce genre de phénomène n'est pas identique au fait de progresser indéfiniment vers la droite du ruban, comme on va le constater bientôt.

Du côté droit du ruban, en effet, après ω_0 cases "normales", se trouve une case "spéciale" présentant le symbole ineffaçable $|\omega_0 \triangleright$ (qui peut être abrégé $|_0 \triangleright$ quand le contexte est suffisamment clair). Une machine de Turing classique n'atteindrait jamais un tel symbole, mais une machine transfinie peut y arriver après ω_0 opérations. L'état dans lequel se trouve $\mathbf{T}_1$ au moment d'aboutir sur $|_0 \triangleright$ dépend des états $q_1, q_2, \dots$, en nombre infini dénombrable, qu'elle a traversés pour arriver jusque-là. Si la séquence $q_1, q_2, \dots$ converge vers un état q , alors nous considérons que la machine est dans l'état q au moment d'atteindre $|_0 \triangleright$. Dans le cas contraire, la machine arrive sur le symbole $|_0 \triangleright$ dans "l'état indéterminé", noté ∞ . Notons que ça ne fait pas d'importance d'envisager la séquence $q_1, q_2, \dots$ à partir de q_1 ou à partir d'un q_n quelconque, puisque la convergence (ou la divergence) de $q_1, q_2, \dots$ est identique à celle de $q_n, q_{n+1}, \dots$. La seule condition, c'est que le point de départ de la séquence soit pris à un instant i postérieur à son dernier passage sur le symbole $|_0 \triangleright$ (s'il y en a eu d'autres). La capacité d'une machine transfinie à déterminer si une séquence infinie de ses états converge ou non est peut-être la plus importante propriété qui la distingue d'une machine de Turing classique.

— Machines stables et instables —

Certaines fonctions f définies sur les nombres entiers naturels sont indéterminées pour des arguments tendant vers l'infini. Un cas évident est la fonction $f(n) = n \bmod 2$. En raison de cette singularité, on peut s'attendre à ce que la réponse d'une machine transfinie capable de calculer des fonctions ne soit pas toujours bien définie. À titre d'exemple, considérons une machine qui écrirait le symbole " p " ou " i " sur la première case de son ruban, selon que le reste du ruban comporte un nombre pair ou, respectivement, impair, de symboles " a ". Sur une entrée présentant un nombre infini de " a ", le résultat de la machine est indéterminé. Sitôt que ω_0 opérations auront été réalisées, on indiquera le symbole "?" sur la première case du ruban. La fonction de transition δ décidera ce qu'elle doit faire lorsqu'elle lit un tel symbole. De façon générale, une machine $\mathbf{T}_\alpha$ est dite *instable* s'il existe au moins une entrée x qui provoque l'apparition d'un ou plusieurs symboles "?" sur la bande, et *stable* dans le cas contraire. De la même manière, une case du ruban est dite stable si le symbole "?" n'y apparaît jamais, et instable s'il existe au moins une façon d'obtenir un tel symbole.

Après que la tête de lecture soit arrivée sur le symbole $|_0\rangle$, trois situations sont envisageables : elle peut soit prendre un état terminal, dans quel cas le programme s'arrête ; soit continuer vers la droite avec l'instruction $\rightarrow$, dans quel cas le ruban se prolonge ; soit revenir vers la gauche avec l'instruction $\leftarrow$. Dans le cas où l'instruction $\rightarrow$ est donnée, le ruban se prolonge d'autant de cases que nécessaire. La machine pourra réaliser un nombre fini d'opérations, ou atteindre le prochain symbole $|_0\rangle$ à l'aide d'un nombre ω_0 d'itérations supplémentaires. Ce nouveau symbole se comportera exactement comme le précédent. Il n'y a d'ailleurs pas de limite au nombre de symboles $|_0\rangle$ que la machine peut traverser, pour peu qu'ils soient en nombre infini dénombrable. Même en se rendant à la position $\omega_0\omega_0$, une machine $\mathbf{T}_1$ trouvera un nouveau symbole $|_0\rangle$. Il n'y a donc aucun moyen pour cette machine d'atteindre le bout de la bande du côté droit. Formellement, on dira que le ruban a pour majorant l'ordinal limite ω_1 , mais qu'en-dessous de cette limite, il peut être arbitrairement grand.

Si la tête de lecture lit un symbole $|_0\rangle$ et que l'instruction $\leftarrow$ lui est donnée, elle revient immédiatement sur le symbole $\triangleright$. Notons que, depuis un emplacement $|_0\rangle$, il serait impossible de réaliser "un seul" déplacement vers la gauche, car en effet $\omega_0 - 1 = \omega_0$ (où " $-$ " dénote la soustraction à gauche). C'est pourquoi l'instruction $\leftarrow$, dans ce contexte, correspond nécessairement à un "retour à la case départ". C'est la seule situation où un saut instantané peut se produire. En termes de complexité temporelle (voir

section 3), on verra plus loin qu'un "retour à la case départ" peut être assimilé indistinctement à un nombre infini dénombrable d'opérations unitaires, à une seule opération $\leftarrow$ ou être vue comme un déplacement instantané, sans qu'aucun de ces choix n'ait d'impact sur la complexité totale⁷.

Ici, on pourrait se demander pourquoi l'instruction $\leftarrow$, à partir d'une case de type $|_0\rangle$, exécute un retour à la case départ, et pas uniquement un "saut" vers le plus proche symbole $|_0\rangle$. Ce choix n'est ni arbitraire ni important. D'une part en effet, il peut être montré que les capacités de calculs qu'on obtiendrait de la part d'une machine munie de cette définition alternative serait équivalentes à celles que l'on retrouve dans notre définition (en ce sens précis que toute fonction calculable ou ensemble semi-décidable dans un cas l'est aussi dans l'autre). Par ailleurs, les deux définitions mèneraient à des complexités temporelles équivalentes, comme il le sera montré plus tard. Pour finir, et il s'agit de la raison décisive, la définition alternative est plus difficilement généralisable à des machines $\mathbf{T}_{\omega_2}, \mathbf{T}_{\omega_3}, \mathbf{T}_{\omega_{\omega_0}}, \dots$ capables d'exécuter plus d'opérations que $\mathbf{T}_{\omega_1}$.

Si nous envisageons $\mathbf{T}_{\omega_n}$, par exemple (avec n un ordinal quelconque), de nouveaux symboles $|_1\rangle, |_2\rangle, \dots, |_{n-1}\rangle$ devront être ajoutés sur la bande, de telle sorte que, pour tout $0 \leq i < n$, les symboles $|_i\rangle$ soient à une distance ω_i les uns des autres. Les règles régissant ces nouveaux symboles sont identiques à ce qu'on a dit plus haut. La machine peut donc y arriver dans un état défini ou dans l'état ∞ , en fonction de la convergence ou de la divergence de la séquence des ω_i états qui ont précédés. Lorsque l'instruction $\leftarrow$ est donnée sur n'importe lequel de ces symboles, elle permet de revenir en $\triangleright$, *et non au précédent symbole de type $|_j\rangle$ présent à gauche* (avec j quelconque). Il n'existe en effet pas toujours de "précédent" symbole $|_j\rangle$. (Pour s'en rendre compte, il suffit de considérer le cas particulier où $\omega_n = \omega_{\omega_0}$, qui est un ordinal fortement inaccessible. Lorsque la tête de lecture est sur le premier symbole $|_{\omega_0}\rangle$ correspondant à cet ordinal, on serait bien en peine de dire quel est le symbole $|_i\rangle$ le plus proche sur la gauche?).

À chaque fois que l'instruction $\leftarrow$ ou $\rightarrow$ est donnée, la fonction de transition peut aussi assigner un nouvel état à la machine, mais ce n'est pas obligatoire. L'état ∞ est toutefois un état spécial, qui n'apparaît que lorsque la séquence d'états ω_0 traversés par la machine est divergente.

Toutes ces remarques nous conduisent à la définition formelle ci-dessous. Par commodité, les notations $\mathbf{T}_{\omega_i}$ et $|_{\omega_i}\rangle$ seront parfois abrégées $\mathbf{T}_i$ et $|_i\rangle$, lorsque le contexte est suffisamment clair. Pour un ordinal α quelconque, on pourra aussi noter $\mathbf{T}_\alpha$ et $|_\alpha\rangle$. Pour finir, on note Σ^* l'ensemble de toutes les chaînes de caractères (finies ou infinies) construites avec un alphabet Σ , en ce compris la chaîne e de longueur nulle.

7. Lorsqu'on définira la notion de complexité temporelle pour des machines transfinies, on verra que la taille des opérations effectuées sera envisagée en termes de cardinalité. Les séquences d'opérations d'ordinalités $\omega_0, \omega_0 + 1, \omega_0 2$ seront donc assignées à une même complexité $\aleph_0$.

Définition 11. Soit un ordinal limite α . Une machine de Turing transfinie T_α est un sextuplet $(K, \Sigma, \delta, s, H, \alpha)$ où

- (a) K est un ensemble fini d'états, comprenant l'état divergent ∞ ;
- (b) Σ est un alphabet contenant le symbole $\sqcup$ (représentant une case vide), les symboles $\triangleright, |_\alpha \triangleright$ (représentant les positions "limites" du ruban) les symboles intermédiaires $|_\beta \triangleright$ (définis pour tout ordinal limite $\beta \subset \alpha$) et le symbole "?" (signalant une indétermination), mais pas les symboles $\rightarrow$ et $\leftarrow$, utilisés exclusivement pour représenter les déplacements de la tête de lecture ;
- (c) $s \in K$ est l'état initial de la machine ;
- (d) $H \subseteq K$ est l'ensemble des états sur lesquels la machine s'arrête ;
- (e) δ est une fonction de transition de $K \setminus H \times \Sigma$ dans $K \setminus \{\infty\} \times \Sigma \cup \{\rightarrow, \leftarrow\}$ telle que :
 1. l'image (q, σ) assigne un état q et une instruction σ du type $\rightarrow$ (déplacement vers la droite), $\leftarrow$ (déplacement vers la gauche) ou $\sigma \in \Sigma$ (qui signifie que la case courante doit être remplacée par le symbole σ) ;
 2. L'ensemble image de δ ne contient pas les symboles "?", $\triangleright, |_\alpha \triangleright$ et $|_\beta \triangleright$ (pour aucun $\beta \subset \alpha$).
 3. Pour tout $p \in K \setminus H$, $\delta(p, \triangleright) = (q, \rightarrow)$.

Pour tout ordinal limite $\beta \subset \alpha$, on observe en outre les règles suivantes :

- (i) lorsque β instructions $\rightarrow$ sont réalisées successivement, alors
 - (i.1) la tête de lecture atteint le prochain symbole $|_\beta \triangleright$ de la bande (en traversant d'éventuels autres symboles de type $|_\gamma \triangleright$, où $\gamma \subset \beta$) ;
 - (i.2) si la séquence d'états $q_0, q_1, \dots, q_\gamma, \dots \in K$ traversés pour arriver sur $|_\beta \triangleright$, depuis le dernier symbole $\triangleright$ ou $|_\theta \triangleright$ rencontré (avec θ quelconque), converge vers un état $q \in K$, alors la machine est dans cet état q , sans quoi elle est dans l'état ∞ ;
 - (i.3) si la séquence de symboles $\sigma_1, \sigma_2, \dots$ se succédant sur une même case au cours de ce même processus converge vers un symbole $\sigma \in \Sigma$, alors la case présente ce symbole, sans quoi elle voit apparaître le symbole "?" ;
- (ii) lorsqu'une instruction $\leftarrow$ est réalisée depuis un emplacement $|_\beta \triangleright$ de la bande, la tête de lecture revient sur le symbole $\triangleright$;
- (iii) la longueur du ruban et le nombre d'opérations totales autorisées par la machine ont α pour majorant^[8].

8. La notation utilisée dans cette définition s'appuie sur la définition d'une machine de Turing de LEWIS R. H. et PAPADIMITRIOU C. H., *Elements of the Theory of Computation*, Prentice-Hall, Upper Sadle River, New Jersey, 1998, définition 4.1.1, p.181.

Par convention, j'ai posé que l'ensemble d'états K soit de taille finie. Il est important de noter qu'une machine qui aurait une infinité d'états présenterait aussi une fonction de transition avec une infinité d'entrées possibles. Cela correspondrait à un algorithme avec une infinité de lignes de code. Il est probable que la puissance de calcul d'une telle machine soit supérieure à une machine de Turing Transfinie, pour peu que la taille de K soit au moins aussi élevée que l'ordinal α utilisé dans la définition de la machine transfinie.

Définition 12. Soit $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ et une chaîne de caractères $x \in \Sigma^*$. On dit que x est une entrée (ou input) de $\mathbf{T}_\alpha$ si et seulement si sa longueur en cardinalité est strictement inférieure à α et qu'elle ne contient pas les symboles $\sqcup$, $\triangleright$ et $|\beta\rangle$ (définis pour tout ordinal limite $\beta \subseteq \alpha$). Par ailleurs, la chaîne de caractères présente sur la bande au moment où $\mathbf{T}_\alpha$ s'arrête (le cas échéant) est appelée une sortie (ou output).

Dans le cas où $\alpha = \omega_1$, la machine $\mathbf{T}_{\omega_1}$ est celle décrite au début de cette section. Une machine $\mathbf{T}_{\omega_2}$, quant à elle, aurait un ruban de cardinalité indénombrable. Sous réserve d'accepter l'hypothèse du continu, on pourrait représenter ce dernier à l'aide d'une droite réelle indexée. À chaque réel r correspondrait un emplacement sur le ruban, qui pourrait être soit blanc $\sqcup$, soit occupé par un symbole de l'alphabet.

À chaque unité de temps, $\mathbf{T}_\alpha$ se trouve dans un état p et lit la lettre σ_{actuel} correspondant à l'emplacement actuel de sa tête de lecture sur le ruban. Si la fonction de transition $\delta(p, \sigma_{\text{actuel}})$ a pour image $(q, \rightarrow)$, la tête de lecture se déplace d'une case vers la droite et la machine passe dans l'état q (dans le cas particulier où $p = q$, il n'y a pas de changement d'état). Si la fonction de transition donne $(q, \leftarrow)$, elle passe dans l'état q et se déplace d'un cran vers la gauche, en "revenant" sur la case départ $\triangleright$ lorsqu'elle exécute cette opération depuis un emplacement $\sigma_{\text{actuel}} = |\theta\rangle$ (pour θ quelconque). Si la fonction de transition a pour image (q, σ) , elle passe à l'état q et remplace la lettre σ_{actuel} , lue par la tête de lecture, par la lettre σ . Dans le cas particulier où $\sigma_{\text{actuel}} = \sigma$, le ruban reste inchangé. Lorsque $\sigma = \sqcup$, la machine efface le symbole lu par la tête de lecture et le remplace par le symbole blanc $\sqcup$.

L'exécution de $\mathbf{T}_\alpha$ se poursuit ainsi tant que le nombre β d'itérations effectuées est majoré par α . Si, à n'importe quel moment, l'état de la machine est dans H , le processus s'arrête. S'il n'existe pas d'ordinal $\beta < \alpha$ tel que cette condition soit satisfaite, alors on dit que la machine boucle " α -infiniment".

Définition 13. Soient deux ordinaux limites $\gamma \subseteq \alpha$, une machine de Turing transfinie $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ et une entrée $x \in \Sigma^*$. Pour chaque ordinal $\beta \subset \alpha$, on note q_β l'état de $\mathbf{T}_\alpha$ après β itérations. On dit que $\mathbf{T}_\alpha$ boucle γ -infiniment sur x s'il n'existe pas d'ordinal $\beta \subset \gamma$ tel que $q_\beta \in H$.

Exemple 2. Considérons la machine de Turing $\mathbf{T}_{\omega_1} = (K, \Sigma, \delta, s, \{h\}, \omega_1)$ où $K = \{q_0, q_1, h\}$, $\Sigma = \{\triangleright, |_0\rangle, \sqcup, a, b\}$, $s = q_0$ et la fonction de transition δ est donnée par la table suivante.

q, σ	$\delta(q, \sigma)$
$q_0 \triangleright$	$(q_0, \rightarrow)$
$q_0 a$	$(q_0, \rightarrow)$
$q_0 b$	$(q_0, \rightarrow)$
$q_0 \sqcup$	(q_1, a)
$q_0 _0\rangle$	$(q_0, \rightarrow)$
$q_1 \triangleright$	$(q_1, \rightarrow)$
$q_1 a$	$(q_1, \rightarrow)$
$q_1 b$	$(q_1, \rightarrow)$
$q_1 \sqcup$	(q_0, b)
$q_1 _0\rangle$	$(h, \rightarrow)$
$\infty \triangleright$	$(h, \rightarrow)$
∞a	$(h, \rightarrow)$
∞b	$(h, \rightarrow)$
$\infty \sqcup$	$(h, \rightarrow)$
$\infty _0\rangle$	$(h, \rightarrow)$

TABLE 10.1

Notons que comme δ a pour domaine $K \setminus H \times \Sigma$, elle doit être définie pour tous les couples (q, σ) possibles. En pratique cependant, les arguments $(\infty, \triangleright)$, $(\infty, \sqcup)$, (∞, a) , (∞, b) ne seront jamais rencontrés (voir la définition 11, point e.2) et peuvent donc être omis de la table (je les ai écrits ici par souci de complétude).

Exemple 3. Considérons la machine de Turing $\mathbf{T}_{\omega_1} = (K, \Sigma, \delta, s, \{h\}, \omega_1)$ où $K = \{q_0, q_1, h\}$, $\Sigma = \{\triangleright, |_0\rangle, \sqcup, a, b\}$, $s = q_0$ et la fonction de transition δ est donnée par la même table que dans l'exemple 2, mais avec l'exception suivante :

q, σ	$\delta(q, \sigma)$
$\infty _0\rangle$	$(q_1, \rightarrow)$

Pour une entrée $x \in \Sigma^*$ de taille quelconque, la machine se comportera de la même façon que dans l'exemple 2, si ce n'est que lorsqu'elle com-

Le programme ci-contre parcourt l'entrée de bout en bout en remplaçant le premier espace blanc $\sqcup$ par un symbole a , le second espace blanc $\sqcup$ par un symbole b ; et en alternant ainsi de façon cyclique. Sur une entrée $x \in \Sigma^*$ de longueur n finie, la machine va lire une infinité d'espaces blancs à partir de l'emplacement $n + 1$. Au moment d'arriver sur le symbole $|_0\rangle$, la suite d'états traversés sera donc constituée d'une succession infinie de $q_0, q_1, q_0, q_1 \dots$. Cette suite diverge, si bien que la machine se trouvera dans l'état ∞ . Comme $\delta(\infty, |_0\rangle) = (h, \rightarrow)$, et puisque h est un état terminal, cela marquera l'arrêt du processus.

Sur une entrée $x \in \Sigma^*$ de longueur infinie n , la machine ne va pas rencontrer d'espace blanc et convergera vers l'état q_0 au moment d'atteindre $|_0\rangle$. Or on sait que $\delta(q_0, |_0\rangle) = (q_0, \rightarrow)$. La machine va donc continuer à arpenter la bande jusqu'à trouver $\sqcup$, en dépassant autant de symboles $|_0\rangle$ que nécessaire. Comme la machine est de type T_{ω_1} , on a $|x| < \omega_1$. Il existera donc forcément un moment où une infinité de $\sqcup$ apparaîtront. La machine va alors poursuivre comme décrit dans le cas de l'entrée finie, et s'arrêter au prochain $|_0\rangle$.

mencera à écrire des $a, b, a, b, \dots$ sur la bande, elle ne s'arrêtera jamais de le faire. En effet, $\mathbf{T}_{\omega_1}$ atteindra le prochain emplacement $|_0\rangle$ dans l'état ∞ . Or $\delta(\infty, |_0\rangle) = (q_1, \rightarrow)$. La machine va donc se déplacer sur la droite et continuer à écrire $a, b, a, b, \dots$

10.2.2 Restrictions et extensions d'une machine transfinie

En quel sens précis peut-on dire de la machine transfinie qu'elle est une extension de la machine de Turing classique ? Dans l'exemple 2 de la section précédente, la machine que nous avons construite boucle ω_0 -infiniment pour toutes les entrées $x \in \Sigma^*$ possibles. Dans le deuxième exemple, en revanche, la machine boucle ω_1 -infiniment (et donc aussi ω_0 -infiniment). Dans un cas comme dans l'autre, les machines exécutent pourtant le "même" programme :

```
programme1(x)
  Se rendre au premier emplacement blanc
  while(true)
    écrire "a"
    écrire "b"
  end
```

En-dessous de $\tau < \omega_0$ étapes, les machines transfinies implémentant le programme ci-dessus se comportent de façons parfaitement indiscernables, mais des différences apparaissent après ω_0 unités de temps. La machine de l'exemple 2 s'arrête en effet, tandis que la seconde continue d'écrire des a et b indéfiniment. Cela démontre que les programmes finis que nous utilisons régulièrement, s'ils étaient donnés à une machine transfinie, seraient ambigus. Il faudrait chaque fois spécifier leurs comportements aux points limites du ruban (ceux qui comportent les symboles $|_\theta\rangle$ pour tous les ordinaux limites $\theta \subset \alpha$).

Définition 14. Soient deux ordinaux limites α et β tels que $\alpha \subset \beta$. On dit que $\mathbf{T}_\beta = (K', \Sigma', \delta', s, H, \beta)$ est une β -extension de $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$, et réciproquement que $\mathbf{T}_\alpha$ est une α -réduction de $\mathbf{T}_\beta$, si et seulement si :

- (a) $K' = K \cup \{\infty\}$;
- (b) $\Sigma' = \Sigma \cup \{|_\gamma\rangle\}$ pour tout ordinal limite $\beta \subseteq \gamma \subseteq \alpha$;
- (c) la fonction δ' est définie de telle sorte que pour tout (q, σ) appartenant au domaine de définition de δ , on ait $\delta'(q, \sigma) = \delta(q, \sigma)$;

Notons que, pour un β fixé, $\mathbf{T}_\beta$ n'est la β -extension que d'une seule machine $\mathbf{T}_\alpha$: celle obtenue à partir de $\mathbf{T}_\beta$ en supprimant toute référence à des symboles $|_\gamma\rangle$ d'ordre supérieur à α . Cependant, l'inverse n'est pas vrai. La machine $\mathbf{T}_\alpha$ a une multiplicité de β -extensions. Cette asymétrie laisse entendre qu'il est possible de connaître le comportement d'une machine réduite

à partir de la machine dont elle est la réduction, mais non le comportement d'une machine étendue à partir de celle dont elle est l'extension.

Définition 15. Soient deux ordinaux $\tau \subset \alpha$ et une chaîne de caractères $x \in \Sigma^*$ de taille égale à β . La troncature en τ de x est la chaîne x obtenue en supprimant tous les caractères de x' au-delà de la position τ .

Définition 16. Soit une machine $\mathbf{T}_\alpha$. La configuration de $\mathbf{T}_\alpha$ a un moment donné est le 3-uple (p, x_1, x_2) , où $p \in K$ représente l'état actuel de la machine, $x_1 \in \Sigma^* \cup \{\triangleright, |_0\triangleright, |_1\triangleright, \dots\}$ est la chaîne de caractères située à gauche de la tête de lecture (y compris la case sur laquelle se trouve celle-ci) et $x_2 \in \Sigma^* \cup \{ |_0\triangleright, |_1\triangleright, \dots\}$ est la chaîne de caractères située à droite de la tête de lecture. Par convention, on écrit un seul $\sqcup$ à partir du moment où le ruban ne se poursuit plus qu'avec des cases blanches.

Exemple 4. Reprenons la machine $\mathbf{T}_{\omega_1}$ de l'exemple 2 et regardons quelle est sa configuration après quatre opérations sur une entrée vide. Elle commence sur $\triangleright$, lit un symbole $\sqcup$, puis écrit successivement a et b . Elle se retrouve maintenant sur une case blanche dans l'état q_1 . La configuration est donc décrite par $(q_1, \triangleright \sqcup a b \sqcup, \sqcup)$.

Proposition 6. Soient une machine $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ et une β -extension $\mathbf{T}_\beta = (K', \Sigma', \delta', s, H, \beta)$. Supposons que $\mathbf{T}_\beta$ soit lancée sur l'entrée $x \in \Sigma^*$ et que $\mathbf{T}_\alpha$ soit lancée sur la troncature en τ de x . Après τ opérations de chaque côté, les deux machines seront dans les mêmes configurations.

Démonstration. Au temps initial, $\mathbf{T}_\alpha$ et $\mathbf{T}_\beta$ commencent avec le même état initial s et les mêmes τ -premiers symboles sur la bande (par la définition 15). Leurs fonctions de transition δ et δ' ont la même image pour tout $p \in K \cap K'$ et $\sigma \in \Sigma \cap \Sigma'$ (définition 14.c). Dès lors, les seuls symboles pour lesquelles des différences peuvent apparaître sont les symboles $|_\alpha\triangleright \in \Sigma'$, les éventuels symboles $|_\gamma\triangleright \in \Sigma'$ (pour tout γ tel que $\alpha \subset \gamma \subset \beta$) et les symboles de x au-delà du τ -ième symbole. Or comme $\tau \subset \alpha$, la machine $\mathbf{T}_\beta$ n'a pas eu le temps d'atteindre aucun de ces symboles.

Remarquons que si $\alpha = \omega_0$, alors les fonctions δ et δ' pourraient aussi différer sur $\in K' \setminus K = \{\infty\}$ (définition 14.a). Cependant, le symbole ∞ ne pourra intervenir que lorsque $\mathbf{T}_\beta$ aura atteint le symbole $|_0\triangleright$, ce qui n'est pas le cas en $\tau \subset \alpha = \omega_0$ opérations. $\square$

Corollaire 1. Soient trois ordinaux limites γ, α, β tels que $\gamma \subseteq \alpha \subset \beta$. La machine $\mathbf{T}_\alpha$ boucle γ -infinitement sur une entrée $x \in \Sigma^*$ si et seulement si sa β -extension $\mathbf{T}_\beta$ boucle γ -infinitement sur x .

Démonstration. On remarque tout d'abord que x est d'ordinalité strictement inférieure à α , sans quoi ce ne serait pas une entrée valide de $\mathbf{T}_\alpha$. Pour tout $\tau \subset \gamma$, $\mathbf{T}_\alpha$ ne s'arrête pas après τ opérations (par la définition

13 de γ -bouclage). Or par la proposition précédente, $\mathbf{T}_\alpha$ et $\mathbf{T}_\beta$ ont la même configuration après τ opérations. Donc pour tout $\tau \subset \gamma$, on conclut que $\mathbf{T}_\alpha$ ne s'arrête pas non plus après τ opérations. La réciproque s'obtient de façon analogue. $\square$

La proposition ci-dessus nous donne la garantie que des expressions comme "la machine $\mathbf{T}_\alpha$ boucle γ -infiniment sur x " ne désignent pas uniquement une propriété de $\mathbf{T}_\alpha$, mais plus généralement, une propriété de l'algorithme implémenté par cette machine. Autrement dit, nous pouvons dire à coup sûr que si un certain algorithme boucle γ -infiniment sur une entrée donnée, alors l'algorithme conserve cette particularité quels que soit la taille du ruban et le nombre d'itérations conférés aux machines l'implémentant.

Une question que l'on pourrait néanmoins se poser est la suivante : existe-t-il des propriétés qui sont conservées quelle que soit la taille de l'entrée x ? En particulier, que se passe-t-il si un algorithme capable de résoudre le problème 3-SAT en temps fini sur une machine de Turing classique est exécuté par une machine $\mathbf{T}_1$ sur une entrée infinie dénombrable ? Va-t-il s'arrêter en temps infini dénombrable ou va-t-il boucler ω_1 -infiniment ? Obtiendra-t-on une réponse à 3-SAT $_{\aleph_0}$? Répondre à ce genre de questions n'est pas du tout évident, car en augmentant la taille de l'entrée, on peut également modifier la sémantique du programme. Pour mettre en exergue cette limitation, nous allons dériver des résultats d'indécidabilité, pour les machines transfinies, analogues au problème de l'arrêt et au théorème de Rice. Mais auparavant, il est nécessaire de se demander quelle est la puissance de calcul des machines transfinies que nous avons construites jusqu'à présent. Quels types de langages sont-elles capables de décider ou de demi-décider ? Peuvent-elles être comparées à des machines de Turing avec oracle ? Et sur cette base-là, est-il possible de construire une machine de Turing transfinie universelle ?

10.2.3 Puissance de calcul

Toute fonction calculable par une machine de Turing classique l'est aussi par une machine de Turing transfinie. Ce fait est établi par la proposition 6. Il en va de même de tous les langages décidables ou semi-décidables. En fait, des différences ne sont susceptibles d'apparaître qu'à partir du moment où au moins ω_0 opérations sont réalisées. On sait que ces différences apparaissent effectivement, en raison de l'existence de fonctions indéterminées (voir l'encadré 10.2.1 sur les machines instables). Par suite, on peut s'attendre à ce qu'un sous-ensemble des langages $\mathcal{L}_\alpha$, composés de mots de longueurs inférieures ou égales à α , soient "incalculables" (dans un sens sans doute différent du problème de l'arrêt). Formellement, on écrira $\mathcal{L}_0 \subseteq \mathcal{D}_\alpha \subseteq \mathcal{L}_\alpha$, où $\mathcal{D}_\alpha$ désigne l'ensemble des langages, constitués de mots de tailles inférieures ou égales à α , qui sont décidables ou semi-décidables par une machine $\mathbf{T}_\alpha$.

Dans l'exemple [3](#) de la première section, on a construit une machine transfinie qui écrit successivement les caractères a et b sur la bande, jusqu'à atteindre le symbole $|_0\rhd$. À chaque étape, cette machine passe de l'état q_1 à q_2 , et ce indéfiniment. Lorsqu'elle atteint le premier symbole $|_0\rhd$, la machine entre dans l'état indéterminé ∞ . On pourrait pourtant faire remarquer que la séquence d'états $q_1, q_2, q_1, q_2, \dots$, bien que divergente, n'en oscille pas moins de façon régulière. Plutôt que d'entrer dans l'état ∞ , la machine pourrait entrer dans un état $\infty_{1,2}$. On parlerait alors de machine k -cyclique, k étant la taille maximale des cycles qu'elle serait capable d'identifier.

Définition 17. Soit un ordinal limite α et $k \in \mathbb{N}$. Une machine de Turing k -cyclique $\mathbf{C}_\alpha = (K', \Sigma', \delta', s, H, \alpha)$ est une machine de Turing transfinie $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ à laquelle on a fait les adjonctions (ou modifications) suivantes :

- K, Σ, s et H sont définis de la même manière que pour une machine de Turing $\mathbf{T}_\alpha$ (voir définition [11](#)).
- Pour tout sous-ensemble d'états $\{q_1, q_2, \dots, q_n\} \in \mathbb{P}(K)$ (avec $2 \leq n \leq k$), il existe dans K' un état cyclique $\infty_{1,2,\dots,n}$ dans K' .
- Pour tout sous-ensemble de caractères $\{q_1, q_2, \dots, q_n\} \in \mathbb{P}(\Sigma)$ (avec $2 \leq n \leq k$), il existe dans Σ' un caractère cyclique $?_{1,2,\dots,n}$ dans Σ' .
- (e) La fonction de transition δ est une fonction de $K' \setminus \{\infty\} \times \Sigma'$ dans $K' \times \Sigma \cap \{\rightarrow, \leftarrow\}$. Il n'existe pas d'image (σ, q) de la fonction δ' telle que q soit un état cyclique ou que σ soit un caractère cyclique. Les règles e.1 à e.3 sont conservées.
- (i.2) si la séquence d'états $q_0, q_1, \dots, q_\gamma, \dots \in K$ traversés pour arriver sur $|_\beta\rhd$, depuis le dernier symbole $\triangleright$ ou $|_\theta\rhd$ rencontré (avec θ quelconque), converge vers un cycle d'états $q_1, q_2, \dots, q_n \in K$ (avec $2 \leq n \leq k$), alors la machine est dans l'état cyclique $\infty_{1,2,\dots,n}$, sans quoi elle est dans l'état ∞ ;
- (i.3) si la séquence de symboles $\sigma_1, \sigma_2, \dots$ se succédant sur une même case au cours de ce processus converge vers un cycle de symboles $\sigma_1, \sigma_2, \dots, \sigma_n \in \Sigma$ (avec $2 \leq n \leq k$), alors la case présente le symbole cyclique $?_{\sigma_1, \sigma_2, \dots, \sigma_n}$, sans quoi elle voit apparaître le symbole "?" ;

Contrairement aux apparences, les langages (semi)-décidés par une machine k -cyclique sont identiques aux langages (semi)-décidés par une machine de Turing transfinie. La raison tient de ce qu'une machine de Turing non cyclique peut écrire sur sa bande des m -uples $(q_1, q_2, \dots, q_m)$ de tailles $2 \leq m \leq k$. À tout instant, ces m -uples peuvent être construits de façon à "garder en mémoire" les m derniers états traversés par la machine de Turing. Si après un nombre infini d'opérations, tous ces m -uples sont remplacés par le symbole "??", cela veut dire que les états successifs $q_1, q_2, \dots$ de $\mathbf{C}_\alpha$ ne convergent pas vers un cycle de longueur m . S'il existe au moins un m -uple qui se stabilise, alors on sait que les états successifs de $\mathbf{C}_\alpha$ convergent vers

un cycle dont la longueur est un diviseur de m . En exploitant cette information, la machine non cyclique peut ainsi reproduire le fonctionnement d'une machine cyclique. Bien sûr, certains problèmes techniques doivent être résolus, comme par exemple : comment doivent être construits ces m -uples ? Où les stocker ? Comment les utiliser ? Le théorème suivant donne les détails de cette approche.

Proposition 7. *Soit une machine k -cyclique $\mathbf{C}_\alpha = (K', \Sigma', \delta', s', H', \alpha)$, avec $k \in \mathbb{N}$. Il existe une machine de Turing $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ à une seule bande qui simule $\mathbf{C}_\alpha$. Plus précisément, supposons que $\mathbf{C}_\alpha$ commence avec $x \in \Sigma^*$ sur son ruban et qu'elle se termine avec $y \in \Sigma^*$. En donnant x comme entrée à $\mathbf{T}_\alpha$, cette dernière termine elle aussi avec y comme sortie. De plus, les temps d'exécution des deux machines sont équipotents.*

Démonstration. On commence par en donner la preuve dans le cas $\alpha = \omega_0$, après quoi on procédera par récurrence. Sans perte de généralité, on peut considérer que $\mathbf{T}_{\omega_1}$ est sous forme normale, si bien que les cases situées en amont du premier symbole $|_0\rangle$ sont vierges. Appelons C_{ω_0+1} la première de ces cases blanches. Appelons de même C_i la i -ème case située directement après le symbole $\triangleright$. La machine $\mathbf{T}_\alpha$ est dotée d'un alphabet $\Sigma = \Sigma' \cup \Sigma'^2 \cup \dots \cup \Sigma'^k \cup \{1, 2, \dots, k\}$, qui lui permet d'écrire des m -uples du type $(\sigma_1, \sigma_2, \dots, \sigma_m)$, avec $\sigma_1, \sigma_2, \dots \in \Sigma'$ et $2 \leq m \leq k$. Notre machine est également capable d'écrire des chiffres sur la bande, allant de 1 à k , et que nous appellerons par la suite "indices de cyclicité". Si un ou plusieurs chiffres entre 1 et k sont déjà inclus dans Σ' , on choisit un encodage de ces chiffres (nous supposons ici, sans perte de généralité, que ce n'est pas le cas). Pour tout état $q \in K$, on suppose également qu'il existe un encodage q^* univoque réalisés avec les lettres de Σ' . Pour chaque état cyclique $\infty_\mu \in K'$ (avec $\mu = (q_1, q_2, \dots, q_m)$) on crée un état q_μ dans K tel que $\delta'(\infty_\mu, \sigma) = \delta(q_\mu, \sigma)$ pour tout $\sigma \in \Sigma'$. On crée de même un état $q_\infty \in K$ tel que $\delta'(\infty, \sigma) = \delta(q_\infty, \sigma)$ pour tout $\sigma \in \Sigma'$. Pour finir, on ajoute à K' des symboles $?_\mu$ tels que, pour tout état $q \in K'$ et pour tout m -uplet $\mu = (\sigma_1, \sigma_2, \dots, \sigma_m)$, on ait $\delta'(q, ?_\mu) = \delta(q, ?_\mu)$.

Avant de commencer la simulation, $\mathbf{T}_\alpha$ présente un ruban vierge. Sur la case C_1 , elle écrit la représentation s'^* de l'état s' dans lequel se trouve initialement la machine $\mathbf{C}_\alpha$. Sur les cases C_2 et C_3 , elle écrit respectivement le nombre 2 et le 2-uplet (s'^*, s'^*) . Sur la case C_4 et C_5 , elle écrit le nombre 3 ainsi qu'un 3-uplet (s'^*, s'^*, s'^*) , et ainsi de suite... Lorsqu'elle arrive aux cases C_{2k-2} et C_{2k-1} , elle écrit le nombre k et le k -uplet $(s'^*, s'^*, \dots, s'^*)$. Sur les cases au-delà de C_{2k} , $\mathbf{T}_\alpha$ va représenter l'entrée x , en suivant une construction analogue.

Notons $\theta \subseteq \omega_0$ la taille de x et $(x_i)_{i < \theta}$ la séquence de caractère constituant x . Pour tout $1 \leq n \leq \theta$, la machine inscrit x_n sur la case C_{2nk} , puis le chiffre 2 sur la case C_{2nk+1} , le 2-uplet (x_n, x_n) sur la case C_{2nk+2} , suivi du nombre 3 et du 3-uplet (x_n, x_n, x_n) , etc. Elle procède ainsi jusqu'à avoir écrit

respectivement le nombre k et le k -uplet $(x_n, x_n, \dots, x_n)$ sur les cases $C_{(n+1)k-2}$ et $C_{(n+1)k-1}$. Le schéma ci-joint représente l'état du ruban à l'état initial.

Notre machine $\mathbf{T}_\alpha$ va ensuite passer dans l'état s' , revenir au début du ruban et simuler le comportement de $\mathbf{C}_\alpha$ de la manière suivante :

- (a) Chaque fois que $\mathbf{C}_\alpha$ lit le caractère écrit sur la case C_n , $\mathbf{T}_\alpha$ lit le caractère écrit sur la case C_{2kn} .
- (b) Chaque fois que $\mathbf{C}_\alpha$ exécute l'opération $\rightarrow$, $\mathbf{T}_\alpha$ fait $2k$ déplacements vers la droite.
- (c) Chaque fois que $\mathbf{C}_\alpha$ exécute l'opération $\leftarrow$, $\mathbf{T}_\alpha$ fait $2k$ déplacements vers la gauche.
- (d) Chaque fois que $\mathbf{C}_\alpha$ écrit un symbole $\sigma \in \Sigma'$ sur la case C_n , $\mathbf{T}_\alpha$ écrit le symbole σ sur la case C_{2nk} . De plus, elle édite les $2k$ cases succédant C_n en suivant les règles ci-dessous. Pour tout $1 \leq m < k$:
 - $\mathbf{T}_\alpha$ lit l'indice de cyclicité i présent sur la case $C_{2nk+2m-2}$ et le remplace par $i + 1$. Si $i + 1 > m$, elle le change par un 1.
 - $\mathbf{T}_\alpha$ se rend sur la case $C_{2nk+2m-1}$ et remplace le m -uplet $(\sigma_1, \sigma_2, \dots, \sigma_m)$ présent sur cette case par un m -uplet identique au précédent, mais où σ_i est remplacé par σ .
 - $\mathbf{T}_\alpha$ revient sur la case C_{2nk} dans le même état que celui qu'elle avait au moment d'y arriver.
- (e) Chaque fois que $\mathbf{C}_\alpha$ passe de l'état p à l'état q (distinct ou non), $\mathbf{T}_\alpha$ passe dans l'état q . De plus, elle va éditer les $2k$ premières cases du ruban en suivant les règles ci-dessous. Pour tout $0 \leq m < k - 1$:
 - $\mathbf{T}_\alpha$ lit l'indice de cyclicité i présent sur la case C_m et le remplace par $i + 1$. Si $i + 1 > m$, elle le change par un 1.
 - $\mathbf{T}_\alpha$ se rend sur la case C_{m+1} et remplace le m -uplet $(q_1^*, q_2^*, \dots, q_m^*)$ présent sur cette case par un m -uplet identique au précédent, mais où q_i^* est remplacé par q^* .
 - $\mathbf{T}_\alpha$ revient ensuite dans l'état q et se rend sur la case où elle se trouvait au moment de passer dans l'état q .
- (f) Si $\mathbf{C}_\alpha$ passe dans un état terminal $h \in H'$, alors $\mathbf{T}_\alpha$ va elle aussi entrer dans un état terminal. Mais auparavant, elle va restructurer le ruban. Le symbole $\sigma \in \Sigma$ écrit sur la case C_{2nk} va être transposé sur la case C_n , et ce pour tout $n \subseteq \theta$.
- (g) Chaque fois que $\mathbf{T}_\alpha$ arrive sur le symbole $|_0\rangle$ dans un état ∞ , elle

fait les choses suivantes :

1. $\mathbf{T}_\alpha$ se rend sur le symbole $\triangleright$ et lit successivement les cases C_{2nk} (pour $1 \leq n \leq \theta$). Si la case C_{2nk} présente le symbole "?", $\mathbf{T}_\alpha$ regarde les cases C_{2nk+1} à $C_{2(n+1)k-1}$ les unes après les autres (par ordre croissant). Si l'une d'elles ne comporte pas le symbole "?", mais un m -uplet $(\sigma_1, \sigma_2, \dots, \sigma_m)$, $\mathbf{T}_\alpha$ rapporte le symbole $_{\sigma_1, \sigma_2, \dots, \sigma_m}$ sur la case C_{2nk} . Sinon, $\mathbf{T}_\alpha$ laisse "?" sur cette case.
2. $\mathbf{T}_\alpha$ se rend ensuite sur la case C_1 . Si cette case comporte le symbole " q^* ", $\mathbf{T}_\alpha$ passe dans l'état q . Sinon, elle lit les cases C_2 à C_{2kn} les unes après les autres (et par ordre croissant). La première case rencontrée qui ne présente pas le symbole "?" fait passer $\mathbf{T}_\alpha$ dans l'état q_μ , où μ est le m -uplet écrit sur cette case. Si toutes les cases C_2 à C_{2kn} présentent le symbole "?", alors $\mathbf{T}_\alpha$ entre dans l'état q_∞ .
3. $\mathbf{T}_\alpha$ effectue l'opération $\rightarrow$ une infinité de fois, sans jamais changer d'état. Elle arrive donc sur le symbole $|_0\rangle$ dans l'état qu'elle avait à la fin de l'étape précédente. Si cet état est q_∞ , alors $\mathbf{T}_\alpha$ fait la même chose que ce que $\mathbf{C}_\alpha$ aurait fait dans l'état ∞ . Si l'état est de type q_μ , où μ est un m -uplet (avec $1 < m < k$), alors $\mathbf{T}_\alpha$ fait la même chose que ce que $\mathbf{C}_\alpha$ aurait fait dans l'état ∞_μ .

Nous ne donnons pas la construction exacte de la fonction de transition de $\mathbf{T}_\alpha$ associée à ces règles, mais nous savons qu'une telle fonction existe. En effet, toutes ces règles sont définies de façon récursive. Elles peuvent donc être implémentées par une machine de Turing classique $\mathbf{T}_0$ judicieusement choisie. Or par la proposition [6](#), $\mathbf{T}_0$ se comporte comme $\mathbf{T}_\alpha$ sur un nombre fini d'étapes. On conclut que $\mathbf{T}_\alpha$ peut implémenter ces règles pour un nombre fini d'étapes arbitrairement grand. Dans le cas où $\mathbf{C}_\alpha$ atteint le symbole $|_0\rangle$, la seule règle concernée est la règle (g) . Cette dernière demande de revenir sur $\triangleright$, ce qu'il est possible de faire avec une machine transfinie à l'aide de l'opération $\leftarrow$. Les autres opérations prescrites par la règle (g) sont définies de façon récursive, et donc exécutables par $\mathbf{T}_\alpha$ pour les mêmes raisons que ci-dessus. On conclut donc que ces règles sont toutes implémentables sur une machine de Turing transfinie.

Il est facile de voir que si $\mathbf{C}_\alpha$ s'arrête sur y en un nombre fini d'étapes, alors la machine $\mathbf{T}_\alpha$ s'arrête aussi sur y . En effet, la règle (f) demande de ne garder que le contenu des cases C_{2kn} . Or en vertu des règles (a) à (d) , ces cases comportent toujours les mêmes symboles que les cases C_n de $\mathbf{C}_\alpha$. Tout ce qu'il reste à montrer, c'est que $\mathbf{T}_\alpha$ peut aussi simuler le comportement de $\mathbf{C}_\alpha$ lorsque cette dernière effectue un nombre infini d'opérations. Le problème, c'est que lorsque $\mathbf{C}_\alpha$ atteint le symbole $|_0\rangle$ dans un état q déterminé (cyclique ou non), $\mathbf{T}_\alpha$, pour sa part, atteint le symbole $|_0\rangle$ dans l'état ∞ . En effet, les règles (d) et (e) , pour être exécutées, demandent à

$\mathbf{T}_\alpha$ de passer dans toutes sortes d'états intermédiaires, si bien que la séquence d'états traversés pour atteindre $|_0\rangle$ ne converge plus. Pour remédier à cette difficulté, la case C_1 de $\mathbf{T}_\alpha$ est construite de telle sorte à représenter l'état de $\mathbf{C}_\alpha$ à n'importe quelle étape du processus (en vertu de la règle (e)). Si les états $q_1, q_2, \dots$ traversés par $\mathbf{C}_\alpha$ convergent vers q , alors les symboles $q_1^*, q_2^*, \dots$ sur la case C_1 se stabiliseront vers la représentation q^* de q . En lisant cette case, $\mathbf{T}_\alpha$ peut donc retrouver l'information comme quoi l'état de $\mathbf{C}_\alpha$ converge bien vers q . C'est d'ailleurs l'esprit des règles (g.2) et (g.3).

Faisons remarquer que si C_1 comporte le symbole "?", cela ne peut vouloir dire que deux choses : (i) soit $\mathbf{C}_\alpha$ est arrivée sur $|_0\rangle$ dans l'état ∞ , (ii) soit $\mathbf{C}_\alpha$ y est arrivé dans un état cyclique ∞_* de taille k . Dans le cas (i), les états $q_1, q_2, \dots$ traversés par $\mathbf{C}_\alpha$ ne convergent vers aucun cycle de longueur k , si bien qu'aucun des m -uples $(q_1^*, q_2^*, \dots)$ obtenus avec la règle (e) ne converge vers un m -uple stable. Par suite, les cases comprises entre C_2 et C_{2kn-1} (inclue) présentent toutes le symbole indéterminé "?". Dans le cas (ii), en revanche, au moins l'un des m -uples présents sur les cases C_2 à C_{2kn} doit s'être stabilisé. Il suffit alors de trouver le plus petit d'entre eux pour connaître le cycle d'états vers lequel $\mathbf{C}$ a convergé, ce qui se produit nécessairement si on parcourt les cases C_2 à C_{2k} par ordre croissant.

Une dernière garantie qu'il faut pouvoir fournir, c'est que les cases C_{2nk} de $\mathbf{T}_\alpha$ restent bien identiques aux cases C_n de $\mathbf{C}_\alpha$ après un nombre infini d'étapes. Lorsque $\mathbf{C}_\alpha$ arrive sur $|_0\rangle$ en effet, certaines cases de son ruban sont susceptibles de présenter un symbole cyclique $?_\mu$ (où μ dénote un m -uple de lettres de Σ'). Comme $\mathbf{T}_\alpha$ n'est pas une machine cyclique, les cases C_{2nk} correspondantes de son ruban comporteront, non pas le symbole $?_\mu$ (inexistant pour $\mathbf{T}_\alpha$), mais le symbole "?". Dès lors, des différences de comportement entre les deux machines risquent d'apparaître. Pour palier à ce problème, la règle (e.2) demande d'écrire, pour toute $m \leq k$, un m -uple $(\sigma_1, \sigma_2, \dots, \sigma_n)$ reprenant les valeurs $\sigma_1, \sigma_2, \dots, \sigma_n \in \Sigma'$ prises par la case C_{2nk} au cours du temps. Si ces valeurs convergent vers un cycle de longueur $l \leq k$, il faut au moins qu'un des m -uples se stabilise. En lisant ce m -uple, la règle (g.1) permet alors de retrouver le symbole $?_\mu$ à indiquer sur la case C_{2kn} .

Toutes ces observations permettent de prouver que $\mathbf{T}_\alpha$ va s'arrêter sur y toutes les fois où $\mathbf{C}_\alpha$ s'arrêtent sur y . Pour vérifier l'équipotence entre les temps d'exécution, il faut faire remarquer que, si $\mathbf{C}_\alpha$ prend un temps infini dénombrable, $\mathbf{T}_\alpha$ prendra aussi un temps infini dénombrable. En effet, chaque fois que $\mathbf{C}_\alpha$ passe par $|_0\rangle$, la machine $\mathbf{T}_\alpha$ passe, au plus, deux fois sur $|_0\rangle_\alpha$ (et au moins une fois). Dans le cas particulier où $\mathbf{C}_\alpha$ s'arrête sur y en un temps fini, ce temps n'est pas équipotent à celui mis par $\mathbf{T}_\alpha$ (car cette-dernière réalise des opérations en plus, à cause des règles (b) à (f)). En revanche, on pourrait construire une autre machine $\mathbf{T}'_\alpha$ qui termine sur y en prenant exactement le même temps. Cette machine serait définie à l'aide du même sextuplet $(K', \Sigma', \delta', s', H', \alpha)$ que la machine $\mathbf{C}_\alpha$ (car sur des temps finis, $\mathbf{C}_\alpha$ se comporte comme une machine non cyclique).

Tout ceci démontre que $\mathbf{C}_\alpha$ peut être simulée par $\mathbf{T}_\alpha$ dans le cas $\alpha = \omega_0$ en un temps équipotents. Pour les cas $\alpha > \omega_0$, on procède par récurrence. Supposons que ce soit vrai pour un certain cardinal $\aleph_n$ quelconque. On va démontrer que ça l'est aussi pour le cardinal $\aleph_{n+1}$ (dans le cadre de cette preuve, nous voyons les cardinaux comme des ordinaux limites). Pour ce faire, on suppose d'abord, sans perte de généralité, que $\mathbf{C}_\alpha$ est sous forme normale. Il n'y a donc que des cases blanches au-delà du symbole $|\aleph_n \triangleright$. Si $\mathbf{C}_\alpha$ prend moins de $\aleph_n$ étapes, l'hypothèse d'induction permet de construire une machine $\mathbf{T}_\alpha$ qui simule $\mathbf{C}_\alpha$. Dans le cas contraire, on peut construire une machine $\mathbf{T}_\alpha$ en suivant les mêmes règles (a) à (g) décrites ci-dessus, mais en prenant $|\aleph_n \triangleright$ au lieu de $|_0 \triangleright$. Nous n'écrivons pas les détails, mais il est facile de voir que cette approche permet de prouver la propriété pour tout $\alpha \subset \omega_0^{\omega_0}$. Dans le cas particulier où α est un ordinal inaccessible, comme par exemple $\omega_0^{\omega_0}$, la preuve par récurrence classique se révèle techniquement insuffisante. Néanmoins, il reste possible de construire une machine $\mathbf{T}_\alpha$, d'après les règles (a) à (g), qui ait le comportement désiré. Alternativement, on peut aussi procéder par récurrence transfinie. □

10.2.4 Représentation minimale

La définition d'une machine $\mathbf{T}_1$ donnée par la définition [11](#) stipule que le ruban peut se prolonger indéfiniment, pour peu qu'il ne dépasse pas en longueur l'ordinal ω_1 . Or on sait qu'un ensemble infini peut être mis en bijection avec une quelconque de ses parties. Par conséquent, il est théoriquement possible de compresser toute l'information du ruban entre le symbole $\triangleright$ et la première occurrence de $|_0 \triangleright$. Supposons par exemple que le ruban contienne une chaîne de $\omega_0^{\omega_0}$ caractères. Cette chaîne peut être vue comme ω_0 séquences ordonnées de tailles ω_0 , et notées $(a_{i,j})_{i \in \mathbb{N}, j \in \mathbb{N}}$ (avec $a_{i,j} \in \Sigma$). Explicitement, on a $a_{0,0}, a_{0,1}, a_{0,2}, \dots, a_{1,0}, a_{1,1}, a_{1,2}, \dots, a_{n,0}, a_{n,1}, a_{n,2}, \dots$. Pour compresser cette information, il suffit de mettre chaque $a_{i,j}$ sur la case donnée par $\pi^2(i, j) = (i + j + 1)(i + j)/2 + j$, la fonction de couplage de Cantor. Cette approche est généralisable à n'importe quel type de ruban, sauf ceux qui sont bornés par un ordinal inaccessible (comme ω_0 ou $\omega_0^{\omega_0}$). Dans ce cas en effet, il n'existe pas d'ordinal $\beta \subset \alpha$ de même cardinalité que α avec lequel construire une bijection. Le ruban d'une machine de Turing classique $\mathbf{T}_0$, en particulier, ne peut pas être mis en bijection avec un ruban de taille finie.

Proposition 8. *Soit un ordinal limite α . Si α n'est pas inaccessible (faiblement ou fortement), appelons β le plus grand ordinal limite tel que $\beta \subset \alpha$. N'importe quel langage décidé ou semi-décidé et n'importe quelle fonction calculée par une machine $\mathbf{T}_\alpha$ peut être décidé, semi-décidé ou calculé (respectivement) par une machine de $\mathbf{T}_\alpha$ qui n'utiliserait que les β premières cases de son ruban.*

Démonstration. La démonstration selon laquelle une machine T utilisant l'ensemble de son ruban peut être simulée par une machine T' , qui n'utilise que les β premières cases de son ruban, découle de la discussion précédente.

Puisqu'il est possible de compresser toute l'information du ruban original sur un ruban de longueur β , il est clair qu'il n'est pas strictement nécessaire d'utiliser des cases supplémentaires. Il suffit de diviser le ruban de longueur β de la machine T' en sous-rubans $R_1, R_2, R_3, \dots$. Le sous-ruban R_1 de T' encode les cases du ruban de T situées entre le symbole $\triangleright$ et le symbole $|\beta\rangle$; le sous-ruban R_2 contient les cases comprises entre les symboles $|\beta\rangle$ et $|\beta_2\rangle$, et ainsi de suite.

Au début de l'exécution, la machine T' se positionne sur la case la plus à gauche du ruban, marquée $\triangleright$. Lorsque les opérations $\rightarrow$ ou $\leftarrow$ sont effectuées, la tête de lecture se déplace vers la case suivante à droite ou à gauche à l'intérieur du sous-ruban R_1 , en sautant toutes les cases appartenant aux autres sous-rubans.

Lorsque T' atteint le symbole $|\beta\rangle$ à la fin des premières β cases, il faut la ramener à la position initiale. Il est nécessaire de mémoriser qu'elle a atteint le symbole $|\beta\rangle$, car à partir de ce point, chaque fois qu'une opération $\rightarrow$ ou $\leftarrow$ est exécutée, la tête de lecture se déplace à la case suivante à droite ou à gauche à l'intérieur du sous-ruban R_2 , sautant toutes les cases appartenant aux autres sous-rubans. Si, à un moment quelconque, la tête de lecture se trouve sur la case initiale $\triangleright$ et que l'opération $\leftarrow$ est exécutée, la machine reste sur le symbole $\triangleright$ mais mémorise qu'elle est revenue au sous-ruban R_1 , plutôt qu'au R_2 .

Pour garder la trace du sous-ruban sur lequel la machine doit opérer, la solution la plus simple consiste à dupliquer chacun des états $q_1, q_2, \dots$ de la machine originale T pour chaque sous-ruban existant. La machine T' sera alors dans l'état $q_{i,j}$ si la machine simulée T serait dans l'état q_i et que la tête de lecture se trouve sur la portion du ruban, telle que représentée dans T' , correspondant au j -ième sous-ruban. $\square$

Forme Normale

La machine $\mathbf{T}_\alpha$ est dite sous *forme normale* si, pour tout cardinal $\beta \subset \alpha$, les cases situées entre la première instance de $|\beta\rangle$ et la première instance de $|\beta+1\rangle$ sont vierges au début de l'exécution de $\mathbf{T}_\alpha$ et le demeurent jusqu'à la fin du processus. La proposition 9 nous donne la garantie que tout langage décidé ou semi-décidé et que toute fonction calculée par une machine $\mathbf{T}_\alpha$ quelconque peut être décidée, semi-décidée ou calculée (respectivement) par une machine $\mathbf{T}'_\alpha$ sous forme normale.

Définition 18. Soient $k \in \mathbb{N}$ et un ordinal limite β . Une machine de Turing minimale $\mathbf{M}_\beta = (K', \Sigma, \delta', s, H, \beta)$ est une machine de Turing transfinie $\mathbf{T}_\beta = (K, \Sigma, \delta, s, H, \beta)$ sous forme normale à laquelle on a fait les adjonctions (ou modifications) suivantes :

- K, Σ, s et H sont définis comme pour une machine de Turing transfinie (voir définition 11).
- K' contient tous les éléments de K ainsi qu'un symbole $\mathbf{1}$ représentant l'état convergent ;
- (e) La fonction de transition δ' est une fonction de $K' \setminus \{\infty, \mathbf{1}\} \times \Sigma'$ dans $K' \times \Sigma \cap \{\rightarrow, \leftarrow\}$. Les règles e.1 à e.3 sont conservées.
- (i.2) si la séquence d'états $q_0, q_1, \dots, q_\gamma, \dots \in K'$ traversés pour arriver sur $|\beta\rangle$, depuis le dernier symbole $\triangleright$ ou $|\theta\rangle$ rencontré (avec θ quelconque), converge vers un état $q \in K'$, alors la machine est dans l'état convergent $\mathbf{1}$, sans quoi elle est dans l'état ∞ ;
- après avoir atteint le symbole $|\beta\rangle$, la machine revient sur la position initiale $\triangleright$.

Proposition 9. Soit une machine de Turing $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \beta)$. Il existe une machine minimale $\mathbf{M}_\beta = (K', \Sigma', \delta', s', H', \beta)$ qui simule $\mathbf{T}_\alpha$. Plus précisément, supposons que $\mathbf{M}_\beta$ commence avec $x \in \Sigma^*$ sur son ruban et qu'elle se termine avec $y \in \Sigma^*$. En donnant x comme entrée à $\mathbf{T}_\alpha$, cette dernière termine elle aussi avec y comme sortie. De plus, les temps d'exécution des deux machines sont équipotents.

Démonstration. Dans la proposition 8, j'ai montré qu'une machine $\mathbf{T}_\alpha$ pouvait être imitée par une machine de Turing $\mathbf{T}_\beta$. Il suffit donc de remarquer que cette machine $\mathbf{T}_\beta$ peut être écrite sous une forme minimale. $\square$

D'un point de vue pratique, notons qu'une machine en représentation minimale doit conserver toujours au moins une case blanche $\sqcup$. En l'absence d'une telle case, son ruban est saturé et il devient impossible d'ajouter des symboles sans en supprimer d'autres. En effet, il n'y a aucun emplacement libre pour mémoriser la valeur d'un symbole qui est transformé par un autre, et cette information est donc perdue irrévocablement (à moins qu'on puisse le retrouver le symbole effacé en effectuant un calcul sur ceux restants on en créant un état spécial dans la machine, mais ce sont des cas particuliers).

Tant qu'il reste une case $\sqcup$ au contraire, la machine peut déplacer tous les symboles d'une case vers la droite. Pour s'en rendre compte, supposons sans perte de généralité que le ruban soit sous la forme $\triangleright \sqcup a_1, a_2, a_3, \dots$, avec $\sqcup$ en première position (un raisonnement tout à fait analogue s'applique si ce n'est pas le cas) et $a_i \in \Sigma$. La machine commence par copier le symbole a_1 sur l'emplacement vide, ce qui donne $\triangleright a_1 a_1, a_2, a_3, \dots$. Ensuite, elle efface le deuxième a_1

10.2.5 Machine à k bandes

À l'aide de raisonnement très analogues à ceux que nous avons produits par une machine minimale, on peut montrer qu'une machine étendue $\mathbf{T}_\alpha^k$, qui posséderait k rubans et k têtes de lecture, n'aurait pas de meilleure puissance de calcul qu'une machine qui n'en posséderait qu'une seule. Pour le montrer, commençons par préciser le fonctionnement de $\mathbf{T}_\alpha^k$.

Définition 19. Soit un ordinal limite α et $k \in \mathbb{N}$. Une machine $\mathbf{T}_\alpha^k$ est un sextuplet $(K, \Sigma, \delta, s, H, \alpha)$, où K, Σ, s et H sont définis de la même manière que pour une machine de Turing $\mathbf{T}_\alpha$ (voir définition 11). La fonction de transition δ est une fonction de $K \setminus H \times \Sigma^k$ dans $K \setminus \{\infty\} \times (\Sigma \cap \{\rightarrow, \leftarrow\})^k$. Pour tout états $p \in K$ et pour tout k -uple $(\sigma_1, \sigma_2, \dots, \sigma_n) \in \Sigma^k$, $\delta(p, (\sigma_1, \sigma_2, \dots, \sigma_n)) = \delta(q, (\sigma'_1, \sigma'_2, \dots, \sigma'_n))$ signifie que la machine passe de l'état q à p et que la i -ème tête de lecture remplace le symbole σ_i de son emplacement actuel par le symbole σ'_i . Par ailleurs, δ respecte les règles (i), (ii) et (iii) de la définition 11, ainsi que les spécificité concernant les symboles spéciaux (11.e.1, 11.e.2 et 11.e3).

La machine n'a donc qu'un seul registre d'état global pour l'ensemble des opérations, mais les têtes de lectures sont capable de réaliser, chacune, une opération indépendante à chaque unité de temps.

Proposition 10. Soit une machine à k -bandes $\mathbf{T}_\alpha^k$, avec $k \in \mathbb{N}$. Il existe une machine de Turing $\mathbf{T}_\alpha$ à une seule bande qui simule $\mathbf{T}_\alpha^k$. Plus précisément, supposons que $\mathbf{T}_\alpha^k$ commence avec $x \in \Sigma^*$ sur la première bande et avec aucun autre caractères sur les $k - 1$ autres bandes. Si $\mathbf{T}_\alpha^k$ termine avec $y \in \Sigma^*$ sur la première bande, alors en donnant x comme entrée à $\mathbf{T}_\alpha$, cette-dernière termine avec y comme sortie.

Schéma de la preuve. Nous savons que la propriété est vraie lorsque $\alpha = \omega_0$, car on se ramène alors au cas des machines de Turing classiques. Nous souhaitons reprendre la même idées de preuve et l'étendre au cas où le ruban est de longueur transfinie.

Nous allons construire une machine $\mathbf{T}_\alpha$ qui simule, sur une seule bande, les processus des k bandes de $\mathbf{T}_\alpha^k$. Pour ce faire, $\mathbf{T}_\alpha$ va découper sa bande infinie en k sous-bande infinies selon une partition échelonnée. Nous représenteront cette partition par une application ϕ de $\mathbb{N} \times \mathbb{O}$ dans $\mathbb{O}$, qui prend en entrée le couple (i, θ) représentant l'emplacement θ de la i -ème bande de $\mathbf{T}_\alpha^k$ et qui renvoie un ordinal θ' correspondant à un emplacement de la bande de $\mathbf{T}_\alpha$. La construction prévoit également d'encoder les positions des têtes de lectures de $\mathbf{T}_\alpha^k$. Pour ce faire, une idée simple consiste à ajouter un symbole μ_i à l'emplacement $\phi(i, \theta)$ lorsque la i -ème tête de lecture de $\mathbf{T}_\alpha^k$ est sur la case θ . Ce symbole μ_k est choisi en dehors de l'alphabet Σ de $\mathbf{T}_\alpha^k$. On étend également l'alphabet Σ' de $\mathbf{T}_\alpha$ en posant $\Sigma = \Sigma' \cup_i (\Sigma \times \mu_i)$. Chaque case θ' de $\mathbf{T}_\alpha$ est donc soit un symbole $\sigma \in \Sigma$, soit un couple (σ, μ_i) .

À chaque étape, $\mathbf{T}_\alpha$ va d'abord simuler le comportement de la première tête de lecture de la première bande de $\mathbf{T}_\alpha$. Pour ce faire, elle va parcourir les emplacements $1, 1k+1, 2k+1, \dots$ correspondant à sa bande jusqu'à trouver un couple (σ, μ_1) . Elle va alors se référer à la fonction de transition δ pour savoir quoi faire. Cette fonction est construite de façon

Pour ce faire, rappelons que $\mathbf{T}_\alpha$ simule $\mathbf{T}_\alpha^k$ en encodant le contenu des k bandes de $\mathbf{T}_\alpha^k$ sur son unique bande. Une manière de faire est donnée par la figure ?? . Elle consiste à étendre l'alphabet Σ' de $\mathbf{T}_\alpha$ en posant $\Sigma' = \Sigma \cup (\Sigma \times \{0, 1\})^k$. Chaque case de la bande de $\mathbf{T}_\alpha$ contient alors un $2k$ -uple de la forme $(a_1, b_1, a_2, b_2, \dots, a_k, b_k)$ avec $a_1, a_2, \dots, a_k \in \Sigma$ et $b_2, \dots, a_k, b_k \in \{0, 1\}$. Ce $2k$ -uple signifie que la première bande de $\mathbf{T}_\alpha^k$ contient le symbole a_1 , que sa deuxième bande contient le symbole a_2 , etc. Par ailleurs, $b_i = 1$ signifie que la tête de lecture de la i -ème bande de $2k$ -uple est située sur la case portant précisément ce symbole a_i . En revanche, $b_i = 0$ signifie que la tête de lecture n'est pas en train de scanner cette case. Les détails de cette approche peuvent être trouvée dans *Elements of the Theory of Computation*⁹.

Nous savons que la propriété est vraie lorsque $\alpha = \omega_0$, car on se ramène alors au cas des machines de Turing classique. Nous allons procéder par induction et supposer, dans un premier temps, que si la propriété est vraie pour un certain β non inaccessible, alors elle est vraie pour le plus petit ordinal limite α tel que $\beta \subset \alpha$. Au vu de la proposition 9, nous pouvons supposer, sans perte de généralité, que $\mathbf{T}_\alpha^k$ n'utilise, pour chacune de ses k bandes, que le tronçon compris entre le symbole $\triangleright$ et $|\alpha\rangle$. Supposons que l'entrée x soit mise sur sa première bande et que, après un certain nombre d'opérations, elle s'achève avec y écrite sur cette bande. Si, à aucun moment du processus, le symbole $|\alpha\rangle$ n'a été atteint, alors la machine s'est comporté exactement une machine de type $\mathbf{T}_\beta^k$ (en particulier, cela veut dire aussi que strictement moins de β caractères de x ont été lus par la première tête de lecture). Par hypothèse d'induction, nous pouvons donc construire une machine $\mathbf{T}_\beta$ qui simule le comportement de $\mathbf{T}_\alpha^k$. Il suffit de lui donner comme input les β caractères de x qui ont effectivement été lus. À partir de là, nous pouvons construire une machine $\mathbf{T}_\alpha$ qui soit l' α -extension de $\mathbf{T}_\beta$. Par la proposition 6, cette machine se comporte exactement comme $\mathbf{T}_\beta$ sur des entrées x de longueur inférieure à β . Elle finira donc par s'arrêter sur la sortie y .

Supposons maintenant que, à une étape τ donnée du processus, une des têtes de lecture de $\mathbf{T}_\alpha^k$ atteigne le symbole $|\alpha\rangle$ après γ opérations. Que la machine atteigne $|\alpha\rangle$ dans l'état ∞ ou dans un état $q \in K$, elle va recevoir un nouvel état $p \in K$ (pas forcément distinct de q) et effectuer un saut vers la gauche (on sait que le saut vers la droite n'a pas lieu car, en nous appuyant sur la proposition 9, nous avons supposé que la machine n'utilisait pas les cases situées derrière l'emplacement $|\alpha\rangle$). La k -ème tête de lecture se trouve

9. LEWIS R. H. et PAPADIMITRIOU C. H., *ibidem*, p.206.

donc maintenant sur le symbole $\triangleright$ dans l'état p . Par hypothèse d'induction, notre machine $\mathbf{T}_\alpha$ peut simuler le comportement de $\mathbf{T}_\alpha^k$ sur toutes les valeurs avant $|\alpha\rangle$, mais au moment où la k -ème bande de $\mathbf{T}_\alpha^k$ atteint $|\alpha\rangle$, $\mathbf{T}_\alpha^k$ atteint elle aussi $|\alpha\rangle$. Il suffit donc d'ajouter à $|\alpha\rangle$ des règles qui vont lui dire $\square$

10.3 État de l'Art

10.3.1 Machine à oracles hiérarchisés

La machine de Turing transfinie que je viens de construire ne serait-elle pas équivalente, en fin de course, à une machine dotée d'un oracle ? Pour rappel, un oracle est une boîte noire, au comportement inconnu, qui est capable de répondre à certaines questions bien précises, comme par exemple : « Cette série infinie converge-t-elle ? » On donne donc une entrée s à cet oracle, potentiellement infinie (une bande contenant des symboles) et on lit la réponse qui est produite (une autre bande contenant des symboles). Si l'oracle est bien construit, il pourrait produire les mêmes symboles que ceux qui sont obtenus par notre machine de Turing transfinie lorsque cette dernière exécute une infinité d'opérations sur sa bande. Tout porte à croire que la machine de Turing Transfinie que j'ai introduite n'est finalement rien qu'un cas particulier d'oracle.

Il existe toutefois deux différences essentielles entre la machine de Turing transfinie et une machine de Turing avec oracles. La première, c'est que l'oracle peut être appelé en une seule unité de temps. Cela permet de décider, quasi instantanément, si une machine M accepte un mot x , alors que cette opération pourrait demander ω_0 étapes de la part d'une machine de Turing transfinie. La seconde, et la plus importante, c'est que l'oracle peut être appelé à n'importe quel moment. Il suffit d'inscrire sur son ruban la "question" qu'on souhaite lui adresser. À l'inverse, une machine de Turing transfinie peut boucler ω_0 -infiniment avant d'atteindre seulement un symbole $|_0\rangle$. De plus, la machine ne permet pas de répondre à n'importe quelle question. La seule information supplémentaire qu'elle fournit, par rapport à une machine de Turing classique, c'est de déterminer si la séquence d'états par laquelle passe la machine de Turing transfinie converge vers un état spécifique. Cette donnée comprend, notamment, la réponse à la question de savoir si la machine s'arrête, car si elle s'arrêtait, elle n'arriverait pas sur le symbole $|_0\rangle$. La machine apprend donc, lorsqu'elle arrive sur $|_0\rangle$, qu'elle ne s'arrête pas — mais elle l'apprend en ω_0 étapes !

Tout se passe en quelque sorte comme si la machine $\mathbf{T}_\alpha$ pouvait passer par différents niveaux de métalangages. Dans le premier, elle se comporte exactement comme une machine classique et ne peut pas avoir recours à aucune forme d'oracle. Dans le deuxième niveau, elle acquiert la possibilité de déterminer l'arrêt de certains types d'opérations réalisées dans le premier

niveau, ce qui lui permettra, comme on va le montrer plus loin, de dépasser le théorème de Rice. Concrètement, cela signifie qu'elle a débloquent l'accès à un oracle permettant de poser toutes sortes de questions sur la sémantique des langages du niveau. De proche en proche, une machine transfinie $\mathbf{T}_\alpha$ va ainsi débloquent une série d'oracles successifs.

On pourrait ainsi être tenté de comparer une machine de Turing transfinie à une machine de Turing classique avec un système d'oracles hiérarchisés. Notons une telle machine $\mathbf{T}'_i$, où i indique le niveau maximum de l'oracle qu'elle est capable de débloquent. L'oracle $\mathcal{O}_0$ prend sur son ruban une entrée x de longueur quelconque (potentiellement infinie), ainsi que l'encodage d'un certain état q , que nous écrirons q^* . Pour distinguer l'encodage de q de la chaîne de caractères x , on peut imaginer qu'un caractère réservé, noté \$, sépare la concaténation de ces deux suites de symboles. $\mathcal{O}_0$ est alors capable de dire si la machine $\mathbf{T}'_i$, lancée sur l'entrée x , converge vers q après un nombre infini dénombrable d'opérations. Formellement, on dira que $\mathcal{O}_0$ décide si x appartient au langage L_q , formé par l'ensemble des mots pour lesquels l'état de $\mathbf{T}_1$ converge vers q . Dans le cas particulier où l'encodage de q est la chaîne de caractères vide, l'oracle $\mathcal{O}_0$ dit si la machine $\mathbf{T}'_i$ s'arrête sur l'entrée x (en temps fini).

Les oracles de niveaux supérieurs sont alors définis de façon récursive. L'oracle $\mathcal{O}_i$ prend en entrée une chaîne $x.\$.q^*$, où le "." dénote la concaténation. Cet oracle est capable de dire si la machine $\mathbf{T}'_i$, lancée sur l'entrée x , converge vers q après un nombre infini dénombrable d'opérations, *et ce sans utiliser les oracles de niveaux strictement supérieurs à i* . Une réponse positive signifie alors que $\mathbf{T}'_i$ converge vers q sur x en utilisant uniquement les oracles $\mathcal{O}_0, \mathcal{O}_1, \dots, \mathcal{O}_{i-1}$. Dans le cas particulier où q^* est la chaîne de caractères vide, $\mathcal{O}_i$ répond à la question de savoir si $\mathbf{T}'_i$ s'arrête sur x (en temps fini) en n'utilisant, comme précédemment, que les oracles de niveaux inférieurs. Dans l'hypothèse où il est possible de simuler une machine $\mathbf{T}_{i-1}$ avec une machine $\mathbf{T}'_{i-1}$, c'est-à-dire, une machine de Turing classique munie des oracles $\mathcal{O}_0, \mathcal{O}_1, \dots, \mathcal{O}_{i-1}$, la question à laquelle répond $\mathcal{O}_i$ consiste à savoir si $\mathbf{T}_i$ boucle ω_{i-1} -infiniment. Nous détenons ainsi une stratégie récursive pour prouver l'équivalence des machines transfinies et des systèmes d'oracles hiérarchisés pour n'importe quel ordinal ω_i .

Pour comprendre comment cette approche fonctionnerait en contexte réel — ou pourquoi elle ne fonctionnerait pas — supposons qu'une machine $\mathbf{T}'_i$ soit conçue de façon à imiter le comportement d'une machine transfinie $\mathbf{T}_i$ donnée. Idéalement, on voudrait que, sur une même entrée x , les deux machines s'arrêtent sur les mêmes sorties y . Par ailleurs, il faudrait mettre en relation biunivoque les temps d'exécution de $\mathbf{T}_i$ et celui de $\mathbf{T}'_i$. Le problème ainsi défini, la première chose que ferait $\mathbf{T}'_i$, lorsqu'elle se lance sur une entrée x , c'est copier son entrée sur l'oracle $\mathcal{O}_0$ en prenant la chaîne de caractères vide comme représentation de q^* . Cet oracle donne une réponse positive si $\mathbf{T}'_1$ s'arrête sur x et négative dans le cas contraire. En cas de

réponse positive, la machine $\mathbf{T}'_1$ fait alors exactement ce qu'aurait fait $\mathbf{T}_{\omega_i}$ sur l'entrée x . Comme $\mathbf{T}_{\omega_i}$ s'arrête après un nombre fini d'opérations, la proposition [6] nous donne la garantie que son comportement puisse être simulé par une machine de Turing classique.

Dans le cas où $\mathcal{O}_0$ répond négativement, en revanche, la machine $\mathbf{T}'_1$ débloque l'oracle $\mathcal{O}_1$ et le lance à nouveau sur l'entrée x , sans aucun encodage pour q^* . En répétant l'opération aussi souvent que nécessaire, elle finit soit par arriver sur un oracle $\mathcal{O}_j$ qui lui répond par la négative, soit par dépasser le niveau i d'oracle maximal autorisé. Dans le premier cas, cela veut dire que la machine boucle ω_{j-1} -infiniment, mais pas ω_j -infiniment. Dans le cas contraire, on en conclut que $\mathbf{T}'_1$ boucle ω_i -infiniment. Cette information est pratique, car elle permet déjà de déduire la complexité temporelle de la machine $\mathbf{T}_1$ simulée par $\mathbf{T}'_1$. Mais le deuxième objectif, rappelons-nous, est de permettre aux deux machines, lorsqu'elles s'arrêtent, de s'arrêter sur la même sortie y . C'est pourquoi $\mathbf{T}'_1$, dès qu'elle atteint un oracle $\mathcal{O}_j$ lui répondant négativement, va simuler les étapes réalisées par $\mathbf{T}_1$ en adoptant la règle suivante : chaque fois que $\mathbf{T}_1$ réalise une opération, $\mathbf{T}'_1$ commence par copier successivement l'état de la bande en cours sur l'oracle $\mathcal{O}_j$ avec, pour encodage de q^* , chacun des états possibles de $\mathbf{T}_1$. En cas de réponse positive à l'une de ces questions, cela veut dire que $\mathbf{T}'_1$ converge vers un certain état q . Dans le cas contraire, cela veut dire que la machine diverge. $\mathbf{T}'_1$ peut en déduire si $\mathbf{T}_1$ va atteindre le symbole $\triangleright_i$ dans l'état q ou dans l'état ∞ . Par commodité, nous supposons que $\mathbf{T}_1$ n'utilise jamais les emplacements situés au-delà du premier $\triangleright_i$, dans l'esprit de la proposition [9].

Il est toutefois difficile, avec cette unique information sur la convergence de l'état de la machine, de simuler exactement le comportement de $\mathbf{T}_1$. Avant d'arriver sur le symbole $\triangleright_i$, en effet, une succession de transformations peuvent être réalisées sur le ruban. Des symboles peuvent être ajoutés, d'autres peuvent être effacés ou remplacés. On aimerait tronquer x à une certaine taille N finie, et décréter que $\mathbf{T}'_1$ est capable de simuler le comportement de $\mathbf{T}_1$ lorsque N tend vers l'infini. Néanmoins, cette stratégie ne garantit pas, en général, l'obtention d'une sortie identique à la troncature de y .

Pour toute taille N fixée, il se peut en effet que des symboles situés au-delà du N -ième emplacement de la bande entraîne une modification des symboles situés en amont, sans pour autant que la sortie soit rendue indéterminée. Un exemple de ce phénomène serait celui d'une machine $\mathbf{T}_1$ ayant les lettres a, b, c pour alphabet et disposant de deux états q_1 et q_2 . Dans l'état q_1 , la machine progresse vers la droite en remplaçant tous les caractères rencontrés par des b , à l'exception des symboles a , qui quant à eux la font entrer dans l'état q_2 . Cet état ordonne à la machine :

1. De mémoriser son emplacement n actuel ;
2. D'effacer tous les caractères situés entre l'emplacement n et le der-

nier symbole a rencontré, qui est lui-même effacé (ou à défaut, d'aller jusqu'au dernier symbole $\triangleright$ ou $|_0\rangle$) ;

3. De revenir sur l'emplacement n et de passer dans l'état q_1 .

Lorsque $\mathbf{T}_1$ arrive sur $|_0\rangle$, elle s'arrête quel que soit son état. Il est facile de voir qu'une telle machine terminera toujours sur $|_0\rangle$ dans l'état q_1 ou q_2 . Par ailleurs, la machine est stable, au sens donné par l'encadré ?? . En revanche, elle est inimitable par un système à oracles hiérarchisés. Quel que soit l'entier N choisi en effet, il est toujours possible qu'un caractère a , voire une infinité, se trouve au-delà de N . Même si l'oracle permet de prédire que la machine va arriver en $|_0\rangle$ dans l'état q_1 , cela ne permet pas de dire qu'aucun caractère a ne va être rencontré en chemin (un nombre pair de a suffit en effet à placer la machine dans l'état q_1).

Au vu de cette discussion, reproduire le comportement de $\mathbf{T}_1$, sur une entrée x infinie, à l'aide d'une machine $\mathbf{T}'_1$ à oracles hiérarchisés, semble impossible en général, bien qu'on puisse le faire sur une troncature de x tendant vers l'infini. Or en vertu de la proposition 6, une machine $\mathbf{T}_1$ exécutée sur une troncature finie a le même comportement qu'une machine de Turing classique. Notre machine de Turing à oracles hiérarchisés se trouve ainsi dans l'incapacité de simuler une machine de Turing transfinie. On pourrait dire, en prenant garde au sens des mots, que l'infini en puissance ne peut pas simuler la capacité de calcul de l'infini en acte, bien qu'une preuve rigoureuse de cette idée mérite encore d'être fournie.

10.3.2 Machine de Turing continue

Une autre question épineuse consiste à se demander si la machine de Turing transfinie que j'ai introduite, en dernier ressort, ne s'apparenterait pas à une machine de Turing continue ? Et, le cas échéant, qu'est-ce que mon formalisme est susceptible d'apporter à ce modèle ?

Une machine de Turing continue, notée $\mathbf{C}$, est une machine analogique. Sa bande de lecture n'est pas divisée en une infinité dénombrable de cases, mais se présente comme une bande magnétique continue, dont on peut lire la valeur en n'importe quel point de l'espace. Formellement, elle est accompagnée d'une fonction $m(t, x)$ associant une valeur à chaque point $x \in \mathbb{R}$ de la bande et à chaque instant t . On présente donc un certain nombre de ressemblances entre cette machine continue $\mathbf{T}$ avec le formalisme des machines de Turing transfinies $\mathbf{T}$ que j'ai précédemment introduit. Je vais commencer par définir plus précisément une machine de Turing continue $\mathbf{C}$. Je dis bien une machine, et non la machine, car il existe plusieurs définitions concurrentielles. Je vais ensuite mettre en relief certains avantages de mon formalisme.

Définition d'une Machine de Turing continue

Pour les besoins de notre modèle, il est supposé que l'espace physique dans lequel notre machine $\mathbf{C}$ est envisagée est un espace continu, dont chaque segment est infiniment divisible. Il est en outre admis qu'on puisse mesurer, avec une précision suffisante, la valeur magnétique en chaque point de cet espace. Ces suppositions peuvent entrer en contradiction avec la mécanique quantique, ou toutefois une certaine interprétation de celle-ci, mais ces réserves théoriques ne nuisent cependant pas au modèle mathématique que nous sommes en train de construire, dont la visée est de capturer la notion de "calcul continu". Je laisse au lecteur le soin de se poser la question de savoir si une telle machine serait "approximable" dans notre monde physique.

Notre machine $\mathbf{C}$, en plus de sa bande continue, dispose d'une tête de lecture. Celle-ci se déplace en suivant une accélération $a(t)$ qui peut être commandée par la machine. Elle ne se rend donc pas d'une case à l'autre, en faisant des sauts de puce, mais se déplace continuellement. Son passage sur la bande, et la lecture du champ magnétique qu'elle fait tout au long de sa trajectoire, peut avoir pour effet de l'accélérer ou de la ralentir. Dans certaines circonstances, cette tête de lecture peut s'immobiliser totalement, dans quel cas on dira que la machine s'est arrêtée. La magnétisation $m(t, x)$ de la bande après l'arrêt définit la sortie s de la machine.

Il est important de remarquer que les états de la machine, eux-aussi, sont représentés par une fonction continue $e(t)$, dont la valeur est susceptible de changer à chaque instant t , en fonction de la valeur magnétique $m(t, x)$ présentée par la bande magnétique à cet instant là, au niveau occupé par la tête de lecture. la fonction accélération $a(t)$ est donc une fonction dépendant de l'état $e(t)$ de la machine, qui a son tour dépend de la position de la tête de lecture et de la magnétisation de la bande.

Par souci de complétude, notons que notre machine a le pouvoir de magnétiser ou de démagnétiser la valeur de la bande $m(t, x + \delta)$, où x est l'emplacement de la tête de lecture et δ est un écart, arbitrairement petit, entre l'endroit où la tête de lecture lit la valeur actuelle de la bande et l'endroit où elle modifie la valeur de la bande (de façon à ce qu'elle ne modifie pas l'état qu'elle est en train de lire au moment où elle est en train de le lire, ce qui pourrait créer une ambiguïté). Il est raisonnable de penser que changer la valeur de la bande en un point affecte également les positions adjacentes. Ces changements pourraient être décrits d'après les équations connues de l'électromagnétisme¹⁰. Dans le cadre d'une machine de Turing continue, on supposera néanmoins que la machine est capable, avec toute la précision requise, de changer la magnétisation au point x sans que cela n'ait

10. Pour une description simple et élégante de cette machine, ainsi qu'un aperçu de certains problèmes soulevés, se référer à l'article de Kevin Brown (<https://www.mathpages.com/home/kmath135/kmath135.htm>).

d'incidence sur les positions adjacentes. C'est une propriété qu'on pourrait appeler la *résolution infinie*.

Cette machine de Turing continue que j'ai décrite s'apparente donc, grosso modo, aux cassettes vidéos analogiques que l'on construisait à l'époque (bien que le programme derrière ces cassettes vidéos soit, quant à lui, très simple : il consiste seulement à lire la bande de gauche à droite à vitesse constante. On pourrait toutefois imaginer un vidéo recoder plus sophistiqué.) Sur le plan purement mathématique, il est important de se demander en quoi la machine de Turing continue dépasse-t-elle la puissance de calcul d'une machine de Turing classique.

Les impasses de la Machine de Turing continue

À première vue, on comprend bien que notre machine **C** dispose de bien plus d'information, sur sa bande magnétique, qu'une machine de Turing classique. Cet avantage en termes de stockage de données n'est susceptible de jouer un rôle que dans la stricte mesure où la machine est capable d'accéder à ces données. Si en effet la machine ne lit pas, au cours de sa lecture, l'ensemble des points de cette bande, ou si cette lecture n'a pas d'impact sur son comportement futur, alors les points non utilisés pourraient, pour le même compte, être supprimés. Il s'agit d'une subtilité d'importance majeure pour notre propos, car il conduit à la question suivante : ajouter du stockage augmente-t-il véritablement les types de calculs qu'une machine **C** peut réaliser, en comparaison avec ce dont était déjà capable une machine de Turing classique ?

Il se trouve que, pour visiter effectivement chacun des points de sa bande, il faut que la machine continue puisse lire et changer le contenu d'un point de sa bande de façon instantanée, chaque fois qu'elle le survole avec sa tête de lecture. Cette propriété est ce que j'appelle la *résolution infinie*. Lorsque par exemple la machine survole un intervalle $[a, b]$ quelconque de la bande magnétique, elle effectue potentiellement une modification de chacun des points survolés. Ces points sont indénombrables, en vertu du théorème de Cantor¹¹. La machine fait donc un nombre infini dénombrable d'opérations à chaque intervalle de temps, aussi petit soit-il.

Admettons la possibilité d'une résolution infinie, concept qui ne pose pas de problème de principe au modèle mathématique, et posons-nous la question suivante : la résolution infinie augmente-t-elle la puissance de calcul de la machine ? Certes notre machine peut lire une fonction réelle en chacun de ces points, mais une fonction réelle n'est rien d'autre qu'une relation entre une entrée et une sortie. On n'a pas besoin de calculer les valeurs de $f(x)$ en chacun de ses points si on sait, par exemple, que $f(x) = x^2$. La valeur

11. Le théorème de Cantor dit que, sur tout intervalle de la droite réel, le nombre de réels compris sur cet intervalle est indénombrable. Il est donc strictement supérieur au nombre d'entiers naturels 1, 2, 3, ...

d'une fonction en chaque point n'est nécessaire que lorsqu'on ne connaît pas la relation unissant les entrées aux sorties. On pourrait donc penser que le pouvoir d'une machine continue ne tient pas tant dans la capacité à calculer des fonctions réelles qu'en la capacité à manipuler des grands ensembles de points en entre lesquelles on ignore les relations fonctionnelles, ou dont ces relations présentent une haute complexité.

Il est possible de donner un sens mathématique à cette notion intuitive de "complexité" en disant qu'il s'agit du nombre minimum de points requis pour décrire la fonction. Une fonction linéaire requiert deux points et est donc très régulière. Une fonction polynomiale d'ordre n requiert $n+1$ points, etc. Une fonction réelle quelconque, pour peu qu'elle soit *continue*, peut-être parfaitement décrite en restreignant le domaine et l'image de la fonction aux rationnels $\mathbb{Q}$. Or les rationnels sont présents en nombre infini dénombrable. Ce que je veux dire par là, c'est que même si une fonction réelle continue est définie sur un ensemble infini indénombrable de points, et qu'elle peut donc être évaluée en un nombre infini indénombrable d'entrées, il suffit de décrire son comportement sur un ensemble infini dénombrables de points pour décrire son comportement entier. Parcourir les nombres rationnels $\mathbb{Q}$, c'est une tâche à la portée d'une machine de Turing classique (même si cette dernière ne peut pas compter tous les rationnels; on reviendra sur cette différence).

Pour contourner cette difficulté, peut être admettra-t-on que la machine de Turing continue soit capable de traiter des fonctions réelles discontinues? Une telle fonction, en effet, ne peut pas être pleinement décrite en regardant uniquement son comportement sur l'ensemble des rationnels. La question qui se pose consiste cependant à se demander combien de discontinuités comporteraient ses fonctions discontinues? Si le nombre de discontinuités est fini, alors cela ne présente pas de bien grande différences avec la situation où notre machine manipulerait des fonctions continues. Il suffirait d'encoder manuellement les points de discontinuité et de les ajouter, pour ainsi dire, à l'ensemble des points rationnels sur lesquels la fonction est définie.

Si maintenant les discontinuités sont en nombre infini, on a véritablement quelque chose de plus qu'avec des fonctions continues. Un exemple célèbre de fonctions discontinues en une infinité de points est la fonction de Dirichlet, qui consiste justement à définir des valeurs différentes sur les points x du domaine appartenant à l'ensemble des rationnels et sur les points x appartenant aux réels. Cette fonction, notée $1_{\mathbb{Q}}$, aussi appelée la *fonction indicatrice*, est définie ainsi :

$$\sigma(s, i) = \begin{cases} 1 & \text{si } x \in \mathbb{Q} \\ 0 & \text{si } x \in \mathbb{R} \end{cases}$$

Ces cas pathologiques mis à part, une machine de Turing continue n'utilise généralement pas plus de pouvoir de calcul qu'une simple machine de

Turing classique. En fait, la différence essentielle tient dans ceci : notre machine de Turing continue a le pouvoir de compter tous les nombres rationnels d'un segment de droite en un temps fini, alors que cela prendrait un temps infini dénombrable ω_0 pour une machine de Turing transfinie. Cette différence est ce que j'appelle la *transcendance*. Elle consiste à atteindre l'au-delà d'un processus énumératif infini, et à manipuler le résultat de ce calcul comme s'il s'agissait d'un objet de pensée à part entière. Or la transcendance n'est-elle pas justement la propriété qui caractérise les machines de Turing transfinie ?

Une machine de Turing Transfinie de type $\mathbf{T}_1$, une fois écrite sous la forme normale, n'est rien d'autre qu'une machine de Turing classique disposant, pour seule propriété supplémentaire, d'accéder "au bout" de sa bande (transcendance). Une machine de type $\mathbf{T}_2$, quant à elle, est une machine capable, en plus de la transcendance, de parcourir une infinité de nombre réels. Il semblerait donc que la machine $\mathbf{T}_2$ de mon formalisme corresponde à la machine de Turing continue. La machine $\mathbf{T}_1$ quant à elle serait une machine continue capable uniquement de parcourir un nombre infini dénombrable de réels. Elle pourrait donc couvrir tous les rationnels et certains réels en proportion négligeable, mais elle aurait, pour ainsi dire, des trous dans sa bande de lecture. Sa précision, quoi qu'élevée, ne correspondrait pas à une *résolution infinie*.

Machine transfinie vs Machine continue

À l'issue de cette discussion, force est d'admettre que la machine $\mathbf{C}$ et la machine transfinie de type $\mathbf{T}_2$ sont très similaires. L'une comme l'autre ont le pouvoir d'énumérer des nombres réels, c'est-à-dire, de compter l'infini indénombrable^[12]. Elles ont également, l'une comme l'autre, la propriété de *transcendance*.

La machine transfinie, il est vrai, prend un temps infini pour "compter les réels" présent sur un segment de droite, tandis que cela prend un temps fini de la part de la machine continue. À ce titre, la machine $\mathbf{C}$ semble plus efficace. Cette distinction quant à la durée de calcul n'est toutefois rien de plus qu'une vue de l'esprit. Il serait tout à fait possible de définir la machine de Turing transfinie de façon à ce qu'elle réalise un comptage infini en un temps fini.

La manière la plus simple de procéder consisterait à dire que le temps qu'elle prend pour réaliser une opération, comme se déplacer d'un cran sur

12. Notons, plus précisément, que $\mathbf{T}_2$ est construite de façon à compter $\aleph_1$ tandis que la $\mathbf{C}$ est capable de compter l'ensemble des réels. Cet ensemble, d'après l'argument de la diagonale de Cantor, est équipotent à l'ensemble des parties des entiers, de cardinalité $2^{\aleph_0}$. La question de savoir si $\aleph_1 = 2^{\aleph_0}$ est une hypothèse connue sous le nom d'hypothèse du continu. Il est aujourd'hui connu, grâce aux travaux de Gödel et de Cohen, que cette hypothèse est indécidable au sein de ZFC. Le système axiomatique le plus couramment adopté aujourd'hui pour construire la théorie des ensembles ne permet donc pas de conclure sur l'existence, ou non, d'un niveau d'infini entre les nombres entiers et les nombres réels.

la bande, lire ou écrire une valeur, décroît au fur et à mesure que la tête de lecture s'éloigne avance le long de la bande. Si l'on note δ_i , la durée d'exécution prise par la machine pour réaliser une opération lorsque sa tête de lecture est sur le i -ème emplacement de la bande (avec $i = 0$ la case initiale), on pourrait poser par exemple :

$$\delta_i = \frac{1}{2^i}$$

Lorsque la machine progresse de gauche à droite, jusqu'à atteindre l'extrémité $\triangleright$ de la bande, elle a donc pris un temps total égal à :

$$\sum_{i=0} \frac{1}{2^i} = 1$$

Cette convention est celle adoptée par le formalisme de la Machine de Zénon¹³. Elle permet à une machine $\mathbf{T}_1$ de "compter tous les nombres entiers" en un temps fini, tout comme le serait une machine de Turing continue. Il est d'ailleurs possible de généraliser cette approche pour permettre à une machine $\mathbf{T}_2$ de compter "tous les nombres réels" en un temps, lui aussi, fini. Pour se faire, il suffit de se rappeler qu'une machine $\mathbf{T}_2$ peut être vue comme disposant d'une bande infinie de taille $\aleph_1$, elle-même décomposée en une infinité de sous-bande infinie de tailles $\aleph_0$. Ces bandes se suivent de façon ordonnée, si bien qu'on pourrait supposer que le temps nécessaire pour faire une opération décroît, aussi bien au sein d'une bande, que lorsque l'on passe d'une bande à l'autre.

La raison pour laquelle je n'ai pas adoptée cette convention de temps de calcul vient de ce qu'elle est inutile. D'un point de vue strictement mathématique, elle ne change rien à la puissance de calcul de la machine. Tout ce qui peut être calculé par l'une en un temps fini peut l'être par l'autre en un temps, au plus, infini dénombrable. Dans un premier cas, on dira que la machine a pris, par exemple, m étapes de calcul, dont trois comptages de l'infini. Dans le second cas, on dira qu'elle en a pris $\omega_0 \cdot 3 + 1$. Je ne vois aucune raison d'exclure cette deuxième écriture. Elle me semble même plus commode.

Pour récapituler, on peut voir la machine de Turing transfinie comme occupant une position intermédiaire entre la machine de Turing classique et la machine de Turing continue disposant d'une résolution infinie. Les avantages de ma description, par rapport à une machine continue sans résolution infinie, sont multiples :

- **Minimalisme** : Une machine de Turing continue $\mathbf{C}$ qui ne dispose pas d'une résolution infinie ne peut pas avoir accès à l'ensemble des

13. Ce concept a été introduit et discuté pour la première fois en 1927 par Hermann Weyl. Une mise en contexte plus contemporain peut être trouvée dans *Zeno machines and hypercomputation* (Petrus H. Potgieter, 2006).

points réels de sa bande de lecture. Ce que je veux dire par là, c'est que bien qu'elle dispose d'une riche information sur sa bande, elle ne pourra manipuler qu'un nombre infini dénombrable de points de la bande. Il est donc plus simple, par un argument similaire au rasoir d'Ockham, d'opter pour ma description des machines de Turing transfinie. Dans cette dernière, la bande ne comprend qu'un nombre infini de points, mais elle peut être parcourue dans toute sa longueur. On a donc éliminé les points qui ne seront pas lus, tout en conservant la propriété vraiment essentielle de la machine.

- **Généralisation :** La machine transfinie $\mathbf{T}_1$ que j'ai introduite peut être généralisée à des machines d'ordre supérieur $\mathbf{T}_2$, $\mathbf{T}_3$, etc. Il se trouve que la généralisation $\mathbf{T}_2$ est la machine $\mathbf{C}$ disposant d'une résolution infinie. Elle possède en effet une bande de lecture dont les cases numérotées vont de zéro à ω_1 . Autrement dit, le nombre de cases est un nombre infini indénombrable. Cette machine peut en outre tourner pendant un temps infini dénombrable, mais aussi pendant un temps infini indénombrable, ce qui permet d'éviter le problème qu'on avait souligné précédemment, concernant le nombre de lignes de code. Une machine $\mathbf{T}_2$ comprend un nombre fini de lignes de codes, mais elle dispose bien d'une résolution infinie.
- **Transposition :** Pour finir, il est facile de saisir, au moyen de mon formalisme, les relations existantes entre une machine de Turing classique, une machine continue à résolution finie et une machine continue à résolution infinie. La propriété 6 dit en effet que, tant qu'une machine de niveau $\mathbf{T}_n$ tourne pendant un temps inférieur à ω_{n-1} , elle se comporte comme une machine de type $\mathbf{T}_{n-1}$. Inversement, si cette machine se met à tourner pendant un temps $\omega n + 1$, elle acquiert la même puissance de calcul qu'une machine de niveau $\mathbf{T}_{n+1}$. Il est donc commode de transposer une machine de Turing classique vers une machine de niveau supérieur, et inversement. Ces machines posséderont les mêmes lignes de codes. Les seules différences possibles concernent des précisions mineures : par exemple, une instruction du type "faire ceci indéfiniment" doit recevoir une précision : indéfiniment pendant un temps infini dénombrable ou indénombrable ? À l'exception de ces transcriptions, qui concernent finalement la précision du langage, la transposition ne demande aucune modification dans le code ou la structure de la machine. Le seul paramètre essentiel, pour le dire autrement, différenciant les machines de Turing classique des machines de Turing continues ou transfinies, c'est le *temps de calcul* (la bande s'étendant aussi loin que la machine est capable d'aller).

Pour achever cette comparaison entre les deux machines, il faudrait en-

core déterminer exactement quelles sont les opérations tombant sous le régime d'une machine de Turing transfinie : est-elle capable de calculer la logique du deuxième ordre ? L'arithmétique du second ordre ? Est-elle sujet à des incomplétudes ou à des indécidabilités ? Cette comparaison en règle dépasserait le cadre de ce travail. Elle demanderait en outre une revue complète de la littérature. C'est pourquoi je me contenterai de montrer, dans les pages qui suivent, des résultats parmi les plus importants : l'existence d'une machine de Turing transfinie universelle, ainsi que l'existence de certains théorèmes d'incalculabilité similaires au problème de l'arrêt introduit par Alan Turing.

10.3.3 Machine de Turing Temporelle Infinie

Dans leur article séminal de 1998, intitulé *Infinite Time Turing Machines*, Joel David Hamkins et Andy Lewis ont introduit un formalisme mathématique apparenté à celui que j'ai développé dans les sections précédentes. Leur approche consiste à étendre la machine de Turing classique de sorte à lui permettre de réaliser une infinité quelconque d'opérations. Une Machine de Turing Temporelle Infinie (MTTI) dotée d'un alphabet $\Sigma = \{0, 1\}$ va donc se comporter comme une machine classique tant que le temps d'exécution reste un nombre fini. Mais dès que le nombre d'opérations effectuées atteint un ordinal limite α donné (α peut être égal à ω_0 , à $\omega_0 2$, etc.), on observe les comportements suivants :

1. La tête de lecture revient dans la position $\triangleright$ initiale du ruban.
2. La machine entre dans un état spécifique q_∞ , qui ne survient que dans ce cas-là.
3. On regarde, pour chaque case c_i , la succession temporelle des symboles écrits sur cette case avant que la machine ait effectué α opérations. Si cette suite tend vers 0, la case c_i présentera le symbole zéro ; dans tous les autres cas de figure, la case présentera le symbole 1.

Équivalence des concepts

Il existe deux différences clefs entre le concept de MTTI ainsi défini et la machine de Turing Transfinie (MTT) que j'ai introduit : (1) tout d'abord, la MTT utilise le symbole ? et non la MTTI ; (2) ensuite, la MTTI possède un ruban de cardinalité $\aleph_0$, tandis que la MTT peut avoir des rubans de cardinalité supérieure.

Le point (1) n'est pas d'une importance capitale. L'absence du symbole ? ne change en effet pas la puissance de calcul de la machine. Pour s'en rendre compte, supposons que nous ayons une machine de Turing Transfinie T_1 et une Machine de Turing Temporellement Infinie T' présentant exactement les mêmes ensembles d'états, les mêmes fonctions de transition, les mêmes alphabets, les mêmes fonctions d'arrêt, etc.

Je vais montrer que toutes les fois où un symbole ? apparaît sur une case c_i de la machine T , il est possible d'obtenir cette information avec la machine T' . Pour s'en rendre compte, il suffit de procéder à une partition de Cantor sur le ruban de T' de telle sorte que chaque case c_i ait un duplicata c'_i . On modifie aussi les instructions de la machine de sorte que, lorsqu'un symbole est écrit sur la case c_i , le symbole inverse soit aussitôt écrit sur la case c'_i . Les cases c_i et c'_i sont donc construites de sorte à présenter le symbole inverse l'une de l'autre.

Dès lors si la case c_i converge vers 1, la case c'_i converge vers 0. Lorsque la case c_i convergera vers 0, la case c'_i convergera vers 1. Et lorsque la case c_i ne converge pas, alors la case c'_i ne convergera également pas, si bien qu'elles présentent toutes les deux le symbole 1. Cela démontre que l'usage du symbole ? n'est pas nécessaire pour déterminer si une case converge ou non.

Le point (2) en revanche est un peu plus fondamental. La MTTI envisagée par Joel David Hamkins et Andy Lewis présente un ruban de taille ω_0 , tandis que les machines transfinies de mon invention ont des rubans de taille potentiellement supérieure. Dit en d'autres termes, Joel David Hamkins et Andy Lewis étendent la mémoire temporelle de la machine de Turing sans toucher à la mémoire spatiale, alors que, de mon côté, j'envisage étendre l'un et l'autre simultanément (et dans les mêmes proportions).

Cette différence souligne que mon approche est plus générale que la leur. La MTTI est conçue uniquement pour réaliser une infinité d'opération sur une bande de longueur infiniment dénombrable. La Machine Transfinie, en revanche, est capable d'opérer un nombre d'opérations infinie dénombrable sur une bande elle-même infinie dénombrable. Les deux concepts se rejoignent néanmoins dans le cas où la cardinalité de la Machine de Turing Transfinie est fixée à $\aleph_0$.

Dans ce cas de figure, en effet, les deux machines ont des comportements tout à fait identiques, si ce n'est que la bande de lecture de la MTTI est d'ordinalité ω_0 , tandis que la bande de la machine de Turing Transfinie se prolonge. Après ω_0 cases, on trouve un nouveau symbole $|\omega_0\rangle$ et ω_0 cases, et ainsi de suite. Cette différence est effacée si l'on représente la machine transfinie dans sa forme normale. Dans ce cas, en effet, seule la première bande de longueur ω_0 est utilisée, et non les suivantes. Cela est assuré par la proposition 9.

On peut donc dire, pour résumer, que la MTTI est équivalente à une machine de Turing Transfinie de cardinalité $\aleph_0$ représentée sous une forme normale.

Comparaison des résultats

Bien que les deux concepts soient équivalents, il peut être utile de faire remarquer que Joel David Hamkins et Andy Lewis, dans leur article, n'abordent

pas les mêmes questions. Le problème qui les intéresse consiste à déterminer, le plus précisément possible, la puissance de calcul d'une MTTI.

L'un de leurs plus grands mérites consiste à avoir introduit la notion d'ordinal atteignable (clockable) et d'ordinal inscriptible (writable). Un ordinal α est atteignable s'il existe une ITTM qui s'arrête exactement au temps α , c'est-à-dire après un nombre d'étapes transfinies égal à α . Un ordinal α est inscriptible s'il existe une ITTM qui, en partant d'un ruban vide, écrit une représentation de α sur son ruban (par exemple, un code binaire de l'ordinal) et s'arrête. À première vue, on pourrait penser que ces deux notions sont équivalentes, et c'est le cas pour tous les ordinaux récurrents. Dès que l'on atteint l'ordinal de Church-Kleene, qui est le plus petit ordinal non récurrent, la situation se corse. Les auteurs démontrent que si tous les ordinaux sont inscriptibles, il existe en revanche des ordinaux qui ne sont pas atteignables.

Une autre contribution de leur article consiste à se demander quels sont les ensembles de réels qui sont décidables ou semi-décidables par une MTTI. Le ruban de la machine est suffisamment large pour contenir un réel, mais pas un ensemble infini indénombrable de réels. Il est toutefois possible de définir un ensemble de réels A au moyen d'une formule du type $x \in A$ si et seulement si $\phi(x)$, où ϕ est une fonction arithmétique. Sous cette approche, la décidabilité consiste à répondre à la question $\forall x \phi(x)$ tandis que la semi-décidabilité consiste à se demander si $\exists x \neg \phi(x)$. Les auteurs démontrent que ces réponses sont possibles avec une MTTI. Ils vont même jusqu'à dire que la machine peut déterminer l'appartenance à des ensembles définies de façon plus complexe, telle que :

$$\begin{aligned} &\forall x \exists y \phi(x, y) \\ &\exists x \forall y \phi(x, y) \end{aligned}$$

En revanche, la MTTI échoue à déterminer des relations qui contiendraient un quantificateur supplémentaire, telles que :

$$\begin{aligned} &\forall x \exists y \forall z \phi(x, y, z) \\ &\exists x \forall y \exists z \phi(x, y, z) \end{aligned}$$

Par comparaison, une Machine de Turing classique ne peut déterminer que des relations sur les nombres naturels qui contiennent un seul quantificateur, et non deux. D'un point de vue formel, on écrira que les MTTI sont capables de décider les ensembles δ_2^1 , tandis que les machines de Turing classique sont limitées à l'ensemble δ_1^0 .

Bien que ces deux résultats soient du plus vif intérêt, ce ne sont pas les questions que j'ai, quant à moi, adressées dans ma démarche. Mon attention

se porte, en effet, non pas sur la puissance de calcul d'une machine transfinie, mais sur les relations exactes existant entre une machine de Turing $\mathbf{T}_0$ (incapable de compter l'infini), une machine de Turing $\mathbf{T}_1$ (capable de compter l'infini dénombrable) et une machine $\mathbf{T}_2$ (capable de compter les réels).

Plus précisément, je me suis demandé si, lorsque l'on construit un algorithme sur une machine $\mathbf{T}_n$ pour résoudre un certain problème sur les réels, on peut transposer cet algorithme sur une machine $\mathbf{T}_{n-1}$ pour résoudre un problème analogue sur les naturels. Si on répond par l'affirmative à ce problème, cela signifie que les deux machines présentent des structures analogues. Dans le cas contraire, cela veut dire que le fait, pour une machine, de compter l'infini indénombrable, change quelque chose de fondamental dans son fonctionnement.

Comme on va le voir, c'est cette deuxième réponse qui est correcte. Pour le montrer, je vais m'intéresser à une version généralisée du problème 3-SAT. Je vais construire un algorithme qui résout ce problème de façon "rapide" pour une machine de Turing Transfinie mais qui, lorsqu'on le transpose à une machine de Turing classique, donne des résultats "lents". Ceci suggère que compter l'infini n'est pas une simple propriété adjonctive, mais change en profondeur la façon dont une machine raisonne. Je tenterai de montrer que cette différence vient de ce que l'infini autorise un mode de raisonnement plus global, plus holistique, en l'absence de quoi une machine se contente d'opérations locales, scindées les unes des autres.

10.4 Machine universelle et problème de l'arrêt généralisé

Je vais donner ici quelques idées pour construire une machine de Turing transfinie universelle. Soit un ordinal $\alpha > \omega_0$ et appelons β le plus grand ordinal limite tel que $\beta \subset \alpha$. Une machine de Turing $\mathbf{T}_\alpha$ sous forme normale contient des parties de son ruban qui sont vierges et qui le demeurent tout au long du processus. Ces parties sont en nombre équipotent à β et sont, chacune, de longueur β . Il est donc possible d'utiliser ces sous-rubans comme des rubans à part entière. Un premier sous-ruban contiendrait l'information permettant de coder le fonctionnement d'une autre machine de Turing $\mathbf{T}'_\alpha$ quelconque. Un autre sous-ruban pourrait être utilisé pour coder l'état de $\mathbf{T}'_\alpha$ à l'instant initial. Un troisième commencerait avec une entrée x de taille supposée inférieure à α . Un quatrième, quant à lui, commencerait vide. À chaque unité de temps, la machine ainsi obtenue va lire le contenu de la troisième sous-bande en considérant qu'il s'agit du contenu lu par la machine $\mathbf{T}'_\alpha$. Utilisant le contenu de la première et de la deuxième sous-bande, notre machine va alors simuler le comportement de $\mathbf{T}'_\alpha$ et en reporter l'input sur la quatrième sous-bande. Cette dernière contiendrait ainsi, à la fin du

processus, la sortie donnée par $\mathbf{T}'_\alpha$ lorsqu'elle est exécutée sur l'entrée x . Une telle machine est notée $\mathbf{U}_\alpha$. Elle va nous permettre de dériver les résultats suivants.

Définition 20. *Soit un ordinal limite β . Le problème de l'arrêt généralisé consiste à décider s'il existe un ordinal limite $\gamma \subseteq \beta$ tel que $\mathbf{T}_\beta$ boucle γ -infiniment.*

Théorème 8 (Théorème de l'arrêt généralisé). *Soient deux ordinaux limites $\alpha \subseteq \beta$. Il n'existe aucune machine de Turing transfinie $\mathbf{T}_\beta$ capable de décider, en un nombre d'étapes $\theta < \alpha$, le problème de l'arrêt généralisé pour une machine $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ et une entrée $x \in \Sigma^*$ quelconques.*

Démonstration. Supposons qu'il existe une machine $\mathbf{T}_\beta$ capable de décider si une machine $\mathbf{T}_\alpha$ quelconque s'arrête en temps $\theta < \alpha$ sur une entrée x quelconque. Notons $\text{halts}(\mathbf{T}_\alpha, x)$ le programme ainsi défini, et $\text{halts}'(\mathbf{T}_\alpha, x)$ la réduction de ce programme sur une machine d'ordinalité α . La proposition 6 nous assure que halts' et halts ont le même comportement pour un nombre $\theta < \alpha$ d'instructions et sur des entrées inférieures à β . En utilisant une approche similaire au théorème d'indécidabilité classique, on pourrait alors construire le programme suivant :

```

diagonal(x)
  if halts'(x, x)
    alors boucle alpha-infiniment
  sinon
    s'arrête
end

```

Notons que l'instruction "boucle β -infiniment" consisterait à envoyer la tête de lecture vers la droite du ruban jusqu'à ce qu'elle atteigne un symbole $|_\alpha\rangle$.

On se demande si $\text{diagonal}(\text{diagonal})$ boucle α -infiniment ? Si ce n'est pas le cas, alors $\text{halts}'(\text{diagonal}, \text{diagonal})$ renvoie *true* et l'instruction "boucle β -infiniment" est lancée, si bien que diagonal boucle α -infiniment. Mais si c'est le cas, alors $\text{halts}'(\text{diagonal}, \text{diagonal})$ renvoie *false* et le programme diagonal s'arrête. Or il ne peut pas s'arrêter en un temps $\theta < \alpha$, sans quoi cela contredirait le fait que $\text{halts}'(\text{diagonal}, \text{diagonal})$ ait terminé sur *false*. Par conséquent, il faut que la fonction $\text{halts}'(\text{diagonal}, \text{diagonal})$, bien qu'elle ait répondu *false*, l'ait fait en un temps $\theta \geq \alpha$.

On conclut qu'il n'existe aucune fonction $\text{halts}'(x, x)$ qui s'arrête en un nombre $\theta < \alpha$ d'étapes. Par le corollaire 1, cela veut dire qu'il n'existe pas non plus de programme halts s'arrêtant en un nombre $\theta < \alpha$ d'étapes. $\square$

Il est également possible de dériver une version généralisée du théorème de Rice.

Définition 21 (Propriété non triviale). *Une machine $\mathbf{T}_\alpha$ possède une propriété P non-triviale si cette propriété est vérifiée pour toutes les entrées possibles de $\mathbf{T}_\alpha$, mais qu'il existe au moins une autre machine de Turing $\mathbf{T}'_\alpha$ pour laquelle au moins une entrée ne respecte pas la propriété P .*

Corollaire 2 (Théorème de Rice généralisé). *Soient deux ordinaux limites $\alpha \subseteq \beta$. Il n'existe aucune machine $\mathbf{T}_\beta$ capable de déterminer en un temps $\theta < \alpha$ si une machine $\mathbf{T}_\alpha$ satisfait une certaine propriété P non triviale sur une entrée x donnée quelconque (une propriété est dite "non-triviale" s'il existe au moins une entrée qui la satisfait et une entrée qui ne la satisfait pas).*

Démonstration. Plan général de la démonstration. Ce corollaire peut être établi en procédant de façon analogue au cas des machines de Turing classiques. Supposons qu'il existe un programme A capable de déterminer, pour un autre programme, s'il possède une certaine propriété non-triviale. On montre alors qu'un tel programme A permettrait de décider, pour n'importe quel autre programme, s'il s'arrête ou non, ce qui contredit l'énoncé du problème de l'arrêt généralisé. $\square$

Définition 22. *Soit un ordinal limite β et un alphabet Σ . Un langage $L \subseteq \Sigma^*$ est β -énumérable (ou β -décidable) s'il existe une machine de Turing $T_\beta = (K, \Sigma, \delta, s, H, \beta)$ qui s'arrête sur tous les mots $l \in L$.*

Une importante conséquence du théorème de l'arrêt généralisé, c'est qu'il existe des langages β -énumérables qui ne sont pas β -décidables par des machines de Turing d'ordinalité $\alpha > \beta$.

Bien qu'une machine universelle transfinie $\mathbf{U}_\alpha$ soit assujettie aux mêmes théorèmes d'indécidabilité que les machines de Turing classiques *concernant l'arrêt des programmes de mêmes ordinalité qu'elle*, on pourrait se demander s'il est possible pour $\mathbf{U}_\alpha$ de déterminer l'arrêt d'une machine d'ordinalité inférieure ?

Pour préciser la question, supposons que nous ayons une machine universelle $\mathbf{U}_\alpha$ et supposons, sans perte de généralité, qu'elle ait quatre sous-bandes à sa disposition. $\mathbf{U}_\alpha$ peut utiliser les trois premières bandes pour simuler la machine $\mathcal{T}_\beta$ donnée sur une entrée x (où $\beta \subset \alpha$ est un ordinal limite et où x est une entrée de taille inférieure à β). La quatrième bande, quant à elle, commencerait vide, mais à chaque fois que $\mathbf{U}_\alpha$ a simulé une nouvelle étape du fonctionnement de $\mathbf{T}_\beta$, on lui demanderait d'ajouter un symbole $a \in \Sigma$ quelconque sur cette bande. De cette manière, la taille de la chaîne de caractères inscrite sur la quatrième bande mesurerait, à chaque instant, le temps d'exécution de $\mathbf{T}_\beta$ (en nombre d'étapes).

Précisons que si la tête de lecture de la quatrième bande atteint un symbole de type $|\theta\rangle$ (pour θ quelconque), il lui est demandé de faire un

mouvement sur la droite et de continuer à écrire des symboles a comme précédemment. Lorsque la simulation de $\mathbf{T}_\beta$ s'achève, il suffit alors de regarder le nombre de symboles sur la quatrième bande pour déterminer combien d'instructions totales ont été réalisées. Ce nombre peut être fini, mais il peut aussi s'agir d'un ordinal $\gamma \subset \alpha$. Dans ce dernier cas, $\mathbf{U}_\alpha$ peut conclure sans faute que la machine $\mathbf{T}_\beta$ qu'elle vient de simuler boucle γ -infiniment sur x . Nous savons en outre que la simulation de $\mathbf{T}_\beta$ par $\mathbf{U}_\alpha$ est correcte, en vertu de la proposition 6.

La machine $\mathbf{U}_\alpha$ que nous venons de décrire serait donc capable de résoudre le problème de l'arrêt pour n'importe quelle machine d'ordinalité β inférieure, mais en prenant, au plus, β itérations. Notons $boucle0(M, x)$ le programme répondant à la question de savoir si une machine M boucle ω_0 -infiniment sur une entrée x finie, et appliquons-la au programme suivant :

```
fonction1(x)
    écrire x sur la bande
end
```

On s'attend à ce que $boucle0(fonction1, "a")$ réponde *false* en un nombre fini d'étapes. Jusque-là, il n'y a rien de surprenant. Cependant, on peut se demander ce qu'il se passe si on ajoute, à la fin de la *fonction1*, l'instruction $boucle0(fonction1, x)$?

```
fonction1(x)
    écrire x sur la bande
    boucle0(fonction1, x)
end
```

En théorie, la nouvelle *fonction1* obtenue devrait, elle aussi, s'arrêter en un nombre fini d'étapes, car l'exécution en série de deux programmes finis s'achève en temps fini. Pourtant, il est facile de voir que *fonction1* va boucler ω_0 -indéfiniment. Dans la mesure, en effet, où la fonction $boucle0(fonction1, x)$ simule *fonction1*, et que cette dernière appelle la fonction *boucle0*, elle va devoir simuler $boucle0(boucle0(fonction1, x), x)$, et ainsi de suite. On en conclut qu'une fonction *boucle0* construite en simulant aveuglément la fonction dont elle est chargée de dire en combien de temps elle s'exécute, est susceptible de modifier ce temps d'exécution. La mesure influence ainsi le résultat de mesure.

Au vu de cette observation, nous pouvons nous demander s'il existe un moyen de construire une fonction $boucle0(M, x)$ qui soustrait de son résultat l'effet de sa présence. En d'autres termes, nous aimerions savoir si une machine de Turing transfinie peut comprendre la sémantique d'un programme d'ordinalité inférieure quelconque autrement qu'en simulant ce-dernier ? (ou avec une méthode plus rapide que la simulation pure et simple). Le théorème

d'indécidabilité généralisée permet justement de répondre à cette question par la négative.

D'autres résultats, plus intéressants, peuvent encore être dérivés.

Proposition 11. ?? Soit un cardinal transfini β . On a les propriétés suivantes :

- (i) Il existe une machine $\mathbf{T}_{\beta+1}$ capable de déterminer, en un temps égal ou inférieur à $\beta + 1$, si une machine $\mathbf{T}_{\beta}$ quelconque boucle β -infinitement sur une entrée x quelconque.
- (ii) Il existe une machine $\mathbf{T}_{\beta+2}$ capable de déterminer, en un temps égal ou inférieure à $\beta + 2$, si une machine $\mathbf{T}_{\beta}$ quelconque boucle β -infinitement sur toutes les entrées x admissibles.
- (iii) Il n'existe pas de machine $\mathbf{T}_{\beta+1}$ capable de déterminer, en un temps égal ou inférieure à $\beta + 1$, si une machine $\mathbf{T}_{\beta}$ quelconque boucle β -infinitement sur toutes les entrées x admissibles.

Démonstration. (i) est obtenu en construisant une machine universelle $\mathbf{U}_{\beta+1}$ sous forme normale qui simule le comportement de $\mathbf{T}_{\beta}$ et utilise une sous-bande vierge pour compter le nombre d'étapes effectuées avant de s'arrêter (ou d'arriver sur le symbole $|_{\beta}$ de la sous-bande).

(ii) et (iii) nécessitent d'écrire $\mathbf{T}_{\beta}$ sous forme normale et de raisonner par l'absurde. On remarque que (iii) faux violerait le théorème de l'arrêt généralisé (la preuve est un peu longue). $\square$

Proposition 12. Soient deux ordinaux $\alpha < \beta$. Soit une propriété P qui soit non triviale aussi bien pour les machines de dimension α que pour celles de dimension β . Le fait qu'une machine $\mathbf{T}_{\alpha}$ possède P n'est pas équivalent au fait que ses β -extension $\mathbf{T}_{\beta}$ possède P .

Démonstration. Supposons par l'absurde que :

1. Si une machine $\mathbf{T}_{\alpha}$ quelconque possède la propriété P non triviale, alors ses β -extension $\mathbf{T}_{\beta}$ possèdent également la propriété P .
2. Si une machine $\mathbf{T}_{\beta}$ quelconque possède la propriété P non triviale, alors son α -restriction $\mathbf{T}_{\alpha}$ possède également la propriété P .

Sous ces hypothèse, montrons qu'on pourrait construire une machine $\mathbf{U}_{\beta}$ capable de déterminer, pour toute machine $\mathbf{T}_{\beta}$, si elle possède la propriété P . Comme cette propriété est non triviale, cela constituerait une violation du théorème de Rice.

Pour construire la machine $\mathbf{U}_{\beta}$, il suffirait d'écrire un algorithme qui vérifie si l' α -réduction $\mathbf{T}_{\alpha}$ de $\mathbf{T}_{\beta}$ possède la propriété P . Cet algorithme consisterait à faire tourner $\mathbf{T}_{\alpha}$ sur toutes les entrées possibles et à vérifier s'il en existe une qui ne vérifie pas la propriété P (c'est possible, puisque $\alpha < \beta$).

Si toutes les entrées vérifient la propriété P , on peut conclure que $\mathbf{T}_\beta$ possède également la propriété P (par l'hypothèse 1). S'il existe au contraire une entrée pour laquelle $\mathbf{T}'_\alpha$ ne vérifie pas la propriété, alors $\mathbf{T}'_\beta$ ne possède pas la propriété P (par la contraposée de l'hypothèse 2). On conclut que les hypothèses 1 et 2 ne sont pas équivalentes.

□

Exemple 5. Soit une machine $\mathbf{T}_{\omega_1}$ qui résout le problème $3\text{-SAT}_{\aleph_0}$ pour n'importe quelle entrée x valide. Une propriété P intéressante consisterait à se demander si $\mathbf{T}_{\omega_1}$ s'arrête après un nombre d'étapes dépendant polynomialement de la taille de l'entrée x . Dans le cas où cette entrée est de longueur ω_0 , cela revient à se demander si $\mathcal{P} = \mathcal{NP}$. Dans le cas où elle est de longueur ω_1 , cela revient à se demander si $3\text{-SAT}_{\aleph_0}$ est soluble en un nombre infini dénombrable d'étapes. Nous allons montrer dans la section suivante qu'il s'agit d'une propriété non triviale. Au vu du théorème ci-dessus, cela veut dire que la solubilité de $3\text{-SAT}_{\aleph_0}$ ne permet pas, en général, de résoudre $\mathcal{P} = \mathcal{NP}$. On peut toutefois contourner cette difficulté en envisageant des relations particulières entre $\mathbf{T}_{\omega_1}$ et sa réduction $\mathbf{T}_{\omega_0}$. C'est l'esprit des remarques qui suivent.

On peut contourner les hypothèses de la proposition [12] de plusieurs manières. Pour commencer, remarquons que cette proposition n'est valide que pour les propriétés P qui sont non triviales quelle que soit la dimension de la machine. Cette condition signifie qu'il doit exister au moins une machine de dimension α qui respecte la propriété P et une machine de dimension β qui ne la possède pas. Mais il faut aussi qu'on ait une machine de dimension β qui respecte la propriété P et une machine de dimension β qui ne la possède pas.

Or dans la pratique, il se peut que la propriété disant qu'un problème soit soluble en un temps polynomial soit une propriété non triviale pour des machines transfinies mais que ce soit une propriété triviale pour les machines de taille finie. Cela serait le cas, notamment, si on avait $P \neq \mathcal{NP}$ dans le cas fini mais $\mathcal{P} = \mathcal{NP}$ dans le cas infini.

Une autre manière de contourner la proposition [12] consiste à envisager qu'on puisse déduire quelque chose à propos des machines infinies en regardant le fonctionnement des machines finies, mais non l'inverse. Cela reviendrait à accepter l'hypothèse 1 figurant dans la démonstration de la proposition [12], tout en rejetant l'hypothèse 2. De façon tout à fait symétrique, on pourrait accepter l'hypothèse 2, mais non l'hypothèse 1 [14].

14. Ma conviction est que cette approche n'est pas viable. Je pense qu'il est possible de poser des conventions, inspirées de la forme normale d'une machine de Turing Transfinie, pour que toute machine de Turing Transfinie de cardinalité α n'admette qu'une, et une seule, β -extension (en guise de justification, remarquons que le formalisme des Time Turing Machine implémente l'idée d'un même algorithme que l'on peut faire tourner en temps

Une dernière approche consiste à se rappeler que, bien que la vérification d'une propriété non triviale P sur des machines transfinies ne permet pas, en général, de conclure quoi que ce soit de significatif à propos de la vérification de cette propriété P sur des machines finies, il peut exister des exceptions. Il n'est pas impossible qu'un certain algorithme donné, ou qu'une certaine classe d'algorithmes, conservent des propriétés lorsqu'on les transpose du cas transfini vers le cas fini.

Il reste donc possible, *a priori*, d'offrir un éclairage à la question $\mathcal{P} = \mathcal{NP}$ à l'aide des machines de Turing transfinies, mais cela demande de trouver une fonction mettant en relation un sous-espace de l'espace des machines de Turing classiques avec un sous-espace de leurs extensions transfinies (c'est l'idée des propriétés P_α et P_β). En fait, on pourrait avoir une injection de l'un vers l'autre, ou de l'autre vers l'un, mais non une bijection.

fini, en temps ω_0 , en temps ω_1 , etc ; j'aimerais faire exactement la même chose, mais en sachant que mes machines à moi présentent des bandes de longueur plus larges). Une fois cela établi, il sera facile de voir que l'hypothèse 1 implique l'hypothèse 2, et inversement, ce qui renforce la proposition 12.

Il faut distinguer deux concepts :

1. La *fonction* est une boîte noire qui associe une réponse à une question, ou un output à un input.
2. Le modèle est une description de la boîte noire, qui en explique le mécanisme interne.

L'abaque, par exemple, est un *modèle* de l'arithmétique. En manipulant l'abaque, on trouve méthodiquement les réponses des *fonctions* « addition » et « soustraction. » Or bien que l'abaque soit un modèle de l'arithmétique, et que l'arithmétique présente des fonctions dont l'esprit humain se rend capable, l'abaque n'est pas pour autant un modèle de l'esprit humain.

De la même manière, l'ordinateur est un *modèle* de toutes les fonctions calculables, mais bien que ces fonctions soient à la portée de l'esprit humain, l'ordinateur n'est pas pour autant un modèle de l'esprit humain. La question qui se pose est de comprendre, dès lors, quelles fonctions non-calculables l'être humain possède en plus de l'ordinateur ?

Dans ce chapitre, j'ai envisagé trois fonctions non-calculables :

1. La complétion de séries infinies.
2. La résolution de problèmes indécidables pour des machines de Turing classique.
3. La résolution rapide de problèmes NP-complets.

J'ai montré que les fonctions (1) et (2) ne pouvaient pas être exécutées par une machine de Turing classique, mais par une machine de Turing transfinie. Dans le prochain chapitre, il s'agira de montrer que c'est aussi le cas de la fonction (3).

Chapitre 11

Problèmes NP-complets généralisés

Dans le chapitre précédent, j'ai défendu que l'intuition pouvait être conçue comme une fonction non calculable, qui posséderait certes une définition explicite, mais qui ne serait pas exécutable par une machine de Turing classique. J'ai en outre envisagé trois façons de construire cette fonction : (1) soit en voyant l'intuition comme une capacité à compter l'infini, ou plus rigoureusement, comme une aptitude à compléter une série infinie ; (2) soit en voyant l'intuition comme la capacité à résoudre des problèmes indécidables, pour lesquels il n'existe pas d'algorithme générique ; (3) soit enfin comme une capacité à trouver, en un temps raisonnable, une solution à un problème NP-complet de taille restreinte.

A priori, ces trois stratégies correspondent à trois visions quelque peu distinctes de l'intuition humaine. La stratégie (1) voit l'intuition comme étant liée à la présence, en nous, de l'idée de l'infini ; c'est parce que nous pouvons manipuler des concepts sur l'infini, et leur donner un sens, que nous ne fonctionnons pas comme des machines. La stratégie (2) voit l'intuition comme l'aptitude à résoudre des problèmes qui résistent au raisonnement logico-déductif ; elle implique donc la présence, dans l'esprit humain, d'une tendance à raisonner autrement que d'après des lois logiques. La stratégie (3) voit dans l'intuition l'exercice d'une sorte de voix intérieure, de guide inconscient, qui nous permet d'avoir « de la chance » face à certaines tâches réputées difficiles. La présence, en nous, de ces idées fulgurantes, ne fait pas objection à la représentation d'après laquelle l'esprit fonctionne comme une machine de Turing. Elle supporte néanmoins l'idée qu'il existe un processus annexe, et pour ainsi dire complémentaire, concernant notre façon de choisir un bon point de départ à notre raisonnement.

J'ai porté ma préférence pour la stratégie (1) et j'ai construit une machine de Turing transfinie. Cette préférence pour (1) vient du fait que l'aptitude à compter l'infini s'apparentait étroitement à la définition de l'intuition

transcendante que j'ai dérivée dans la théorie tricamérale de l'esprit (voir section [8.4](#)). Pour rappel, l'intuition transcendante est en effet l'aptitude à comprendre qu'un concept tel qu'un « arbre » dissimule une infinité d'arbres possibles, s'apparentant et se différenciant partiellement les uns des autres. Le concept « d'arbre » est conçu comme la limite d'un processus de comparaison où chaque terme du langage renvoie à une multitude d'autres termes. Il y a donc, d'après moi, un lien entre l'intuition de l'infini et la construction de concepts abstraits universels.

Construire une machine satisfaisant la propriété (1) n'était pas chose aisée. À la sueur de mon front de mathématicien, j'ai étendu le pouvoir de calcul de la machine de Turing classique de manière à lui permettre d'opérer une infinité d'opérations. Concrètement, je lui ai permis de se rendre au « bout » du ruban qui lui est attribué, de façon à lire le « dernier » terme d'une série infinie. Bien sûr, j'ai défini le « bout » de la bande et le « dernier » terme en m'inspirant de la théorie des infinis de Cantor, sans quoi ces expressions n'auraient aucun sens. Cela m'a conduit à construire, non pas une seule machine de Turing transfinie, mais une infinité de machines $\mathbf{T}_{\omega_1}, \mathbf{T}_{\omega_2}, \dots$. La machine $\mathbf{T}_{\omega_1}$, aussi appelée machine d'ordre 1, est capable d'énumérer l'infinité des nombres entiers, la machine $\mathbf{T}_{\omega_2}$ peut compter l'infinité des nombres réels, la machine $\mathbf{T}_{\omega_2}$ peut compter tous les ensembles infinis possibles de nombres réels, etc.

À la façon d'un architecte qui, après avoir jeté les fondations d'une tour sur le papier, se rend sur le terrain pour étudier le dénivelé, j'ai cherché à étudier quelques-unes des propriétés concrètes à propos de ce nouveau formalisme mathématique. J'ai, en particulier, cherché à me demander s'il existait des problèmes indécidables pour une machine de Turing transfinie. À cette fin, je suis parti du problème de l'arrêt conçu par Alan Turing, défini de la manière suivante : « Existe-t-il un algorithme permettant de déterminer si une machine de Turing T quelconque s'arrête en un temps fini sur une entrée s donnée ? » J'ai montré qu'il était possible de construire une machine de Turing transfinie permettant de résoudre le problème de l'arrêt pour une machine de Turing classique. En revanche, il n'existe pas d'algorithme permettant à une machine de Turing transfinie de déterminer si une autre machine transfinie, d'ordre équivalent ou supérieur, s'arrête sur une entrée s . Autrement dit, le problème de l'arrêt est généralisable aux machines de Turing transfinies.

De ces considérations, il s'ensuit qu'il existe une hiérarchie infinie de problèmes indécidables. Certains problèmes sont indécidables pour des machines de Turing classiques, d'autres le sont pour des machines de Turing transfinies d'ordre 1, puis pour des machines transfinies d'ordre 2, etc. Par suite de quoi mon formalisme des machines transfinies permet aussi bien de compter l'infini que de résoudre certaines classes de problèmes indécidables. Autrement dit, mon outil, conçu initialement pour la stratégie (1), constitue également une bonne approche pour la stratégie (2). Ce qui signifie,

en dernier ressort, que la conception d'après laquelle l'intuition est la présence, en nous, d'une idée de l'infini, coïncide avec l'idée que nous puissions résoudre, par intuition, des tâches insolubles au moyen d'une machine de Turing classique.

Dans ce chapitre, j'aimerais développer deux questions qui sont laissées en suspens. La première consiste à savoir si les machines de Turing transfinies peuvent aussi nous apprendre quelque chose au sujet de la stratégie (3). Autrement dit, je me demande si une machine transfinie est en mesure de trouver une solution rapide à un problème NP-complet, réputé pour être fastidieux à résoudre. Si tel est le cas, cela signifie que les trois visions de l'intuition humaine que nous avons identifiées coïncident étroitement. En termes techniques, je dirai qu'il s'agit de trois fonctions coextensives. Bien qu'elles puissent être représentées séparément, le modèle des machines transfinies, conçu pour exprimer l'une d'entre elles, supporte également les deux autres. Ce sont les trois facettes d'un même fonctionnement plus général.

Par suite de quoi, et il s'agit de la deuxième question que je voudrais aborder dans ce chapitre, il se pourrait que l'intuition holistique, dont j'ai parlé dans la théorie tricamérale de l'esprit, soit également coextensive des autres définitions de l'intuition données précédemment. Pour rappel, cette définition stipulait que l'esprit humain est capable d'appréhender l'information de façon globale, en la décomposant d'après des structures perceptives déterminées, plutôt que de la traiter comme une succession d'unités binaires attendant à être liées les unes aux autres.

Pour exprimer cette thèse sous une forme lapidaire, il se pourrait que l'intuition, cette hydre aux têtes multiples, ne soit qu'un seul et même monstre multiforme ; qu'il existe une connexion conceptuelle forte entre la conception de l'infini, la compréhension holistique de la perception que nous avons vue dans les illusions d'optique et la résolution de certains problèmes indécidables ou complexes — tels que les problèmes NP-complets.

Définitions	Fonctions
Intuition transcendante	Compléter des séries infinies Solutionner des problèmes indécidables
Intuition holistique	Solutionner des problèmes NP-complets Sujet aux illusions perceptives

Dans les pages qui vont suivre, je vais d'abord rappeler la définition des problèmes NP-complets et montrer en quoi la résolution de ces problèmes exige généralement une durée exponentielle. Je vais ensuite montrer que les machines de Turing transfinies ne sont pas plus efficaces, sur ces problèmes, qu'une machine de Turing classique. La raison en est qu'une machine de Turing transfinie, lorsqu'on lui demande de réaliser un nombre fini d'étapes de raisonnement, se comporte en tout point identiquement à une machine

de Turing classique.

En revanche, je vais construire une généralisation des problèmes NP-complets qui est telle que, pour une machine transfinie, trouver une solution à ces problèmes n'est pas plus difficile que de vérifier si une réponse donnée est une solution. Ce problème NP-complet généralisé dont il est question n'est rien d'autre qu'un problème NP-complet faisant intervenir des fonctions continues sur la droite réelle. C'est donc un problème défini sur $\mathbb{R}$ et non sur $\mathbb{N}$. Ce problème comporte une infinité de variables, mais il correspond à des problèmes de géométrie dont on peut se faire une représentation intuitive. Si cette notion de problèmes NP-complet généralisé peut paraître encore flou aux yeux du lecteur, j'entends la clarifier dans les sections qui vont suivre.

11.1 Les Problèmes NP-complets classiques

Les problèmes NP-complets, comme on l'a dit dans la section 6.6, forment une seule et même classe de structure équivalente. Il faut comprendre par là que tout algorithme de résolution permettant de solutionner un problème P_1 de cette classe peut, moyennant une adaptation simple, être utilisé pour résoudre un autre problème P_2 de cette même classe. Par « adaptation simple », il est entendu que la transformation de l'algorithme se fait en temps polynomial ; ou autrement dit, que le temps de calcul requis pour changer l'algorithme n'est pas significatif compte tenu du temps qui est déjà nécessaire à sa résolution proprement dite.

Cette importante propriété fait qu'il est possible de choisir un problème de la classe des problèmes NP-complets pour les représenter tous. Ce choix se porte généralement sur le problème 3-SAT, qui consiste à résoudre une formule logique. Par souci de commodité, cette formule peut être représentée sous forme normale conjonctive. Cette représentation consiste à écrire la formule sous la forme d'un ensemble de clauses du type « A ou B ». On peut penser par exemple aux clauses « A ou non B », « non A ou B », « non B ou non A », ce qui donne la formule F suivante :

$$(A \vee \neg B) \wedge (\neg A \vee B) \wedge (\neg B \vee \neg A)$$

Il est facile de vérifier que cette formule admet une seule solution, à savoir A et B faux tous les deux. Si en effet A est vrai, alors la troisième clause nous dit que B est faux. Mais si B est faux, la deuxième clause implique que A doit être faux. On conclut par l'absurde que A doit être faux. De façon symétrique, supposons que B soit vrai. La troisième clause nous enseigne que A est faux. Or si A est faux, la première clause implique que B doit être faux, ce qui contredit la supposition que B était vrai. Il s'ensuit que B ne peut être que faux. Ou autrement dit, il faut leur assigner la valeur $A = B = 0$.

La confirmation de ce raisonnement est donnée dans la table de vérité ci-dessous. Pour rappel, $\wedge$ et $\vee$ sont les quantificateurs logiques "Et" et "Ou" respectivement, tandis que $\neg$ représente la négation :

A	B	$(A \vee \neg B)$	$(\neg A \vee B)$	$(\neg B \vee \neg A)$	F
1	1	1	1	0	0
1	0	1	0	1	0
0	1	0	1	1	0
0	0	1	1	1	1

Ce problème est un problème NP-complet, ce qui signifie que vérifier si une assignation des variables A et B est une solution peut se faire rapidement, tandis que trouver la bonne assignation des variables est susceptible de prendre plus de temps. Plus précisément, on dira que le nombre d'étapes requise pour vérifier que $B = B = 0$ est une solution du problème demande de vérifier chaque clause une à une. Il existe donc une relation linéaire entre le nombre de clauses et le temps d'exécution requis pour la tâche. En revanche, trouver la solution $A = B = 0$, si on ne s'y prend pas adéquatement, exigera de parcourir l'ensemble des solutions possibles et de les vérifier une à une. En présence de deux variables A et B , il n'y a que quatre possibilités :

- $A = B = 1$
- $A = 1$ et $B = 0$
- $A = 0$ et $B = 1$
- $A = B = 0$

Si le problème comportait trois variables A, B et C , il aurait fallu huit étapes à notre raisonnement. S'il comportait quatre variables, il en aurait fallu seize, etc. En présence de n variables $v_1, v_2, \dots, v_n$, il faut 2^n opérations. En revanche, vérifier qu'un candidat solution identifiée satisfait effectivement les clauses ne demande jamais que n opérations. Il suffit en effet de parcourir les clauses une à une. On remarque sans peine que 2^n est une expression exponentielle, tandis que n est une expression polynomiale. Il en découle la définition du problème NP-complet :

Définition 23. *Un problème NP-complet est un problème dont la vérification d'une solution peut être réalisée en un temps polynomial tandis que la découverte d'une solution est, au moins, polynomiale, et au maximum exponentielle.*

Les problèmes NP-complets sont omniprésents : lorsque vous tentez de résoudre le jeu du démineur, que vous mettez des objets dans votre sac de randonnée, que vous répartissez des tâches avant un souper de famille, que vous formez des binômes d'étudiants en tenant compte de leur préférence,

que vous cherchez à optimiser un itinéraire passant par un certain nombre de magasins, ou chaque fois que vous tentez de trouver une solution d'un problème pouvant s'écrire sous la forme d'une équation logique, vous faites face à un problème NP-complet.

Personne ne sait s'il existe une méthode de résolution générale permettant de trouver la solution d'un problème NP-complet en un temps polynomial. Toutes les méthodes connues actuellement présentent des faiblesses ; si bien que, quelle que soit la robustesse de la méthode, il existe des situations où elle exigera approximativement 2^n étapes. En des termes techniques, on dira qu'on ignore si les problèmes de type NP-complets appartiennent, ou non, à la classe P des problèmes solubles en temps polynomial. On écrira $\mathcal{P} = \mathcal{NP}$ si tel est le cas, et $P \neq NP$ dans le cas contraire.

L'expérience suggère que, face à ce genre de problèmes, l'esprit humain ne procède pas toujours au moyen d'une logique rigoureuse, mais en tentant sa chance. Lorsqu'il s'agit par exemple de jouer au Démineur, vous réalisez rapidement que la décision de cliquer sur une case précise du jeu demande d'analyser les cases adjacentes et, de proche en proche, l'ensemble du plateau de jeu. Il existe une relation globale entre le choix de cliquer sur la case supérieure gauche et les informations dont vous disposez sur des cases pourtant très éloignées. Notre traitement de l'information étant limité, nous sommes amenés à poser des hypothèses raisonnables, à reconnaître des schémas récurrents, etc. Le record du monde est de 1 seconde pour le niveau débutant (tableau de 8x8 cases), 8 secondes pour le niveau intermédiaire et 29 (16x16 cases) secondes pour le niveau expert (31x16 cases). Il va sans dire qu'une telle rapidité ne peut pas être obtenue en mobilisant uniquement l'analyse logique (le système 1 de Kanheman), qui est lente, froide, séquentielle.

La question que l'on se pose, dès lors, c'est si l'intuition peut être caractérisée comme une certaine aptitude à identifier rapidement un bon candidat solution à ce genre de problème. Dans la mesure où la définition formelle que j'ai donnée de l'intuition est la machine de Turing transfinie, la question revient à se demander dans quelle mesure cette machine présenterait un avantage, pour résoudre un problème NP-complet, par rapport à la machine de Turing classique. S'il était démontré, en effet, que le recours à l'infini permettrait de mieux comprendre les liens globaux existant entre les variables du problème, de sorte à accélérer le processus de résolution, l'on détiendrait un argument montrant qu'il y a bel et bien une connexion entre les deux définitions de l'intuition envisagées jusqu'à présent : celle du tout et celle de l'infini.

Hélas, il ne faut pas beaucoup d'efforts pour se rendre compte que ce n'est pas le cas. Une machine de Turing transfinie ne diffère, par rapport à la machine de Turing classique, que lorsqu'on lui donne un temps de réflexion infini. Mais tant que le nombre d'étapes réalisées par la machine transfinie reste fini, la proposition [6](#) nous enseigne que cette machine transfinie ne peut pas être distinguée d'une machine de Turing classique. Cela signifie que l'une

comme l'autre ont la même efficacité algorithmique.

Ce résultat compromet-il irrévocablement la démarche ? Pas nécessairement. Il est vrai que la machine de Turing classique et la machine de Turing transfinie ont la même efficacité face à des problèmes NP-complets de tailles finies. Mais dans la mesure où la machine de Turing transfinie est une extension infinie de la machine de Turing classique, la vraie question pertinente consiste à se demander comment se comporte cette machine face à des extensions infinies du problème NP-complet ? Jusqu'à présent, en effet, nous nous sommes limités à des problèmes possédant un nombre fini de variables $v_1, v_2, \dots, v_n$. Or si ce nombre tendait vers l'infini, quelle efficacité aurait nos algorithmes ?

Pour une machine de Turing classique, on s'attend logiquement à ce que la vérification d'une solution tende vers $\lim_{n \rightarrow \infty} n$, tandis que la recherche d'une solution, avec les algorithmes actuellement connus, tendrait vers $\lim_{n \rightarrow \infty} 2^n$, si bien que le rapport entre les deux donnerait :

$$\lim_{n \rightarrow \infty} \frac{2^n}{n} = 0 \quad (11.1)$$

Ce résultat, si l'on prend la peine d'y songer, est sensiblement logique. Dans la mesure où le nombre de variables tend vers l'infini, il faut un temps infini pour vérifier qu'une solution satisfasse le problème. La recherche d'une solution, en revanche, au vu des algorithmes connus, consistera à vérifier toutes les solutions possibles. Il faut donc vérifier, une infinité de fois, des solutions dont la vérification, elle-même, prend un temps infini. Dans la théorie de George Cantor, on dira que la vérification de la solution prend un temps infini dénombrable, noté $\aleph_0$, tandis que la découverte d'une solution prend un temps infini non dénombrable donné par $2^{\aleph_0}$. On comprend bien que $2^{\aleph_0}$ est un infini plus grand que $\aleph_0$.

Dans les sections qui vont suivre, je vais démontrer un théorème contre-intuitif : pour les machines de Turing transfinies, il est possible de construire un algorithme de résolution tel que si le problème NP-complet comporte effectivement une infinité de variables, alors le rapport donné par l'équation 11.1 vaut 1. Autrement dit, il n'est pas beaucoup plus difficile, pour une machine de Turing transfinie, de résoudre un problème NP-complet que de vérifier qu'une réponse constitue bien une solution au problème. Pour le dire encore autrement, les résultats que je vais exposer suggèrent que la classe NP^T des problèmes NP-complets transfinis est équivalente à la classe P^T de problèmes polynomiaux transfinis. On peut donc écrire $NP^T = P^T$.

11.2 Le Problème 3-SAT_{N₀}

Pour étudier l'efficacité algorithmique des machines de Turing transfinies, il convient tout d'abord de généraliser les problèmes NP-complets à un nombre infini de variables. Le problème NP-complet le plus connu, on

l'a dit, est le problème 3-SAT, qui consiste à résoudre une équation logique. Je vais généraliser ce problème 3-SAT à un nombre infini dénombrable de variables.

Considérons un ensemble infini dénombrable de $\eta \subset \omega_1$ **variables propositionnelles D** (avec $\eta \subset \omega_1$). Dans ce qui va suivre, nous les numéroterons sans perte de généralité $v_1, v_2, \dots$. Cela ne veut pas dire que $\eta = \omega_0$, mais que les variables peuvent être mises en bijection avec les entiers naturels. Considérons également un ensemble infini dénombrable de $\mu \subset \omega_1$ **clauses** de la forme $l_i \vee l_j \vee l_k$ (avec i, j, k des entiers naturels distincts). Chaque **littéral** l_n figurant dans cette expression peut soit dénoter la variable propositionnelle v_n (dans quel cas on parle de littéral positif), soit son contraire $\neg v_n$ (littéral négatif). Le problème **3-SAT** $_{\aleph_0}$ consiste à trouver une assignation des variables propositionnelles qui satisfasse l'ensemble des clauses.

Nous allons montrer que ce problème est soluble à l'aide d'une méthode utilisant, au plus, un nombre infini dénombrable d'étapes. Pour ce faire, une représentation géométrique de 3-SAT $_{\aleph_0}$ va d'abord être avancée, dans laquelle le domaine des solutions admissibles sera assimilé au segment de droite $[0, 1]$. Ensuite, les clauses vont être exprimées au moyen **d'harmoniques** Ψ s'annulant sur certaines valeurs clefs de $[0, 1]$. Pour finir, un algorithme de résolution va être introduit. Par souci d'efficacité de calcul, cet algorithme sera écrit sous une forme indépendante de la représentation géométrique donnée plus haut, mais celle-ci servira pour prouver la convergence de l'algorithme avec le plus de clarté et d'élégance possible. En matière de conclusion, nous montrerons que l'algorithme, écrit sous forme de pseudo-code, peut être implémenté par une **machine de Turing transfinie** T_{ω_1} telle que définie dans la section [10.2](#).

11.3 Représentation géométrique du problème

Il est possible de donner à 3-SAT $_{\aleph_0}$ une représentation géométrique. Au lieu d'affecter les valeurs *true* ou *false* aux variables propositionnelles $v_1, v_2, \dots \in V$, nous pourrions leur donner une valuation ν de D dans $\{0, 1\}$. On dira que $v_i \in D$ est vraie si et seulement si $\nu(v_i) = 1$, et fausse dans le cas où $\nu(v_i) = 0$. L'espace des solutions admissibles de 3-SAT $_{\aleph_0}$ est alors un espace de Cantor : $\{0, 1\}^{\aleph^*}$. Chaque élément de cet espace peut être assimilé à l'expansion binaire d'un nombre réel de l'intervalle $[0, 1]$. Il suffit pour ce faire de poser une application ϕ qui associe la séquence $(v_i)_{i \in \aleph^*}$ au nombre $\sum_{i=1}^{\infty} \nu(v_i)/2^i$. La séquence "101010...", par exemple, serait représentée par $2/3$ et "010101..." par $1/3$.

Dans cette représentation, chercher une solution de 3-SAT $_{\aleph_0}$ revient à trouver un réel sur le segment de droite $[0, 1]$ qui satisfait les clauses $l_i \vee l_j \vee l_k$ (avec i, j, k des entiers naturels distincts). Or on sait depuis Cantor que la

cardinalité de n'importe quel segment de la droite réelle est $2^{\aleph_0}$. Par suite, une machine capable de réaliser $\aleph_0$ opérations ne pourrait pas balayer naïvement l'espace des solutions admissibles, ni même des sous-intervalles de ce segment (à l'exception des singletons de type $[a, a]$, qui ne comprennent qu'un seul point). À moins qu'il existe une infinité indénombrable de solutions au problème, la probabilité de trouver, par chance, une bonne réponse, est strictement égale à zéro¹.

Un lecteur rigoureux fera remarquer que l'application ϕ posée plus haut n'est pas une injection de $\{0, 1\}^{\mathbb{N}^*}$ dans $[0, 1]$. En effet, les séquences se terminant par une infinité de 1 ou de 0 ont la même image. Par exemple, "0111..." et "1000..." ont toutes les deux $1/2$ pour image, mais ne désignent pas les mêmes séquences, ni d'ailleurs les mêmes *limites de séquences*. Supposons en effet que nous construisions une suite infinie dénombrable de séquences $v_0, v_1, \dots$ telle que $\phi(v_i) < 1/2$ (pour tout $i \in \mathbb{N}$) et $\lim_{n \rightarrow \infty} \phi(v_n) = 1/2$. Quelle que soit la façon dont nous les choisissons, les séquences de cette suite s'approcheront tendanciellement de la séquence "0111...", qui commence par 0 et suivie d'un nombre toujours plus grand de 1 consécutifs. À l'inverse, si nous construisons une suite $v_0, v_1, \dots$ telle que $\phi(v_i) > 1/2$ et que $\lim_{n \rightarrow \infty} \phi(v_n) = 1/2$, alors la suite tendra vers une séquence "1000..." initiée par un 1 et suivie d'un nombre toujours plus grand de zéros consécutifs. En des termes encore intuitifs, on dira qu'un point appartenant à l'ensemble image de ϕ peut être approché à gauche et à droite de façon différente. Cette singularité ne se produit cependant que pour des rationnels bien spécifiques. Or comme les rationnels sont négligeables dans l'ensemble des réels, on pourrait dire que ϕ est injective presque partout sur le segment de droite $[0, 1]$.

Sur cette parenthèse méthodologique, la question que nous nous posons est celle de savoir si les clauses de 3-SAT $_{\aleph_0}$ sont assez riches en information pour nous permettre de trouver une solution en un nombre, au plus, infini dénombrable d'étapes. Plutôt que de chercher les assignations de $\nu(v_i)$ qui satisfont l'ensemble des contraintes, on pourrait se demander quelles sont celles qui les falsifient. Parfois en effet, le problème dual est plus facile à se représenter que le problème de départ. En utilisant la loi de Morgan, chaque clause $l_i \vee l_j \vee l_k$ est équivalente à $\neg(\neg l_i \wedge \neg l_j \wedge \neg l_k)$. Une clause ainsi réécrite exclut de l'espace des solutions admissibles certains sous-intervalles de $[0, 1]$. La clause $\neg(\neg v_1 \wedge \neg v_2 \wedge \neg v_3)$, par exemple, dit en substance que la solution ne peut pas être dans l'intervalle $[0, 1/8)$. En fait, le point $1/8$ est exclu si $(v_i)_{i \in \mathbb{N}} = 000111\dots$ et inclus si $(v_i)_{i \in \mathbb{N}} = 001000\dots$. Cette observation suggère la définition suivante.

Définition 24. Soient deux réels $a, b \in [0, 1]$, une séquence $v \in \{0, 1\}^{\mathbb{N}^*}$ et la fonction $\phi : \{0, 1\}^{\mathbb{N}^*} \rightarrow \mathbb{R} : (v_i)_{i \in \mathbb{N}^*} \rightarrow \sum_{i=1}^{\infty} \nu(v_i)/2^i$. Un intervalle

1. Rappelons que, en présence d'ensemble de cardinalités infinies, des événements de probabilité nulle ne sont pas forcément impossible mathématiquement.

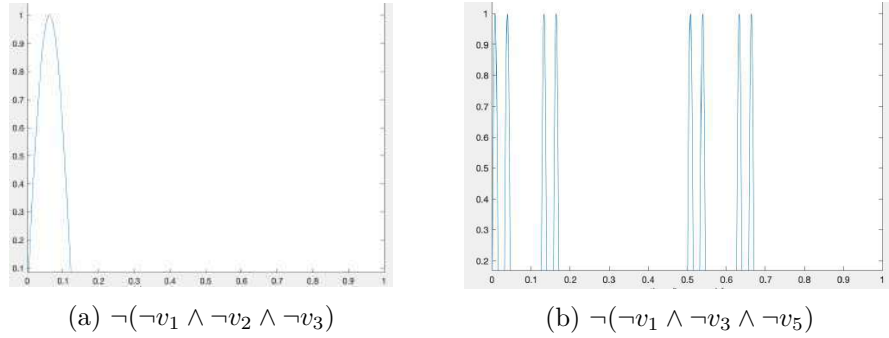


FIGURE 11.1 – Représentation graphique des intervalles exclus par deux clauses différentes. L'axe horizontal représente le segment réel $[0, 1]$, correspondant à l'ensemble des solutions possibles. À chaque clause est associée une fonction périodique $\psi : [0, 1] \rightarrow \mathbb{R}$ telle que, si $x \in [0, 1]$ est exclu par la clause, alors soit $\psi(x) > 0$, soit $\psi(x) = 0$ et la dérivée $\psi'(x) > 0$. L'axe vertical indique la valeur de cette fonction périodique.

d'exclusion, noté $|a, b|$, dénote un sous-ensemble de $\text{Dom}(\phi)$ tel que $v \in |a, b|$ si et seulement si une des trois conditions ci-dessous est respectée :

1. $\phi(v) \in (a, b)$.
2. $\phi(v) = a$ et v ne se termine pas par une infinité de 1. On dira alors que a est un point limite gauche de l'intervalle.
3. $\phi(v) = b$ et v ne se termine pas par une infinité de 0. On dira alors que b est un point limite droit de l'intervalle.

Exemple 6. La clause $\neg(\neg v_1 \wedge v_2 \wedge \neg v_3)$ a pour intervalle d'exclusion $|2/8, 3/8|$. On dit de $2/8$ et $3/8$ que ce sont des points limites gauche et droit (respectivement). La mesure de cet intervalle est $1/8$ (voir figure 11.1a).

Exemple 7. La clause $\neg(\neg v_1 \wedge \neg v_3 \wedge \neg v_5)$ a pour intervalle d'exclusion $|0, 1/32| \cup |2/32, 3/32| \cup |8/32, 9/32| \cup |9/32, 10/32| \cup |1/2, 17/32| \cup |18/32, 19/32| \cup |24/32, 25/32| \cup |26/32, 27/32|$. Cet intervalle est la réunion de huit intervalles disjoints de mesure $1/32$ (voir figure 11.1b).

Proposition 13. Pour tout réels a, b, c et d , les trois expressions ci-dessous sont équivalentes :

1. $|a, b| \cup |b, c| = |a, c|$;
2. $|0, a| = [0, a]$;
3. $|a, 1| = [a, 1]$.

Démonstration. Une séquence $v \in \{0, 1\}^{\mathbb{N}^*}$ satisfait $\phi(v) = 0$ si et seulement si elle ne se compose que de 0. Elle satisfait $\phi(v) = 1$ si et seulement si elle ne se compose que de 1. Cela montre les points 1 et 2.

Pour le point 1, supposons que $v \in \{0, 1\}^{\mathbb{N}^*}$ satisfasse $\phi(v) = b$. Dans ce cas, soit v ne se termine pas par une infinité de 0 et alors $\phi(v) \in |a, b|$, soit v ne se termine pas par une infinité de 1 et alors $\phi(v) \in |b, c|$ (il s'agit d'une disjonction inclusive). Dans les deux cas, $\phi(v) \in |a, c|$. Les situations où $\phi(v) \neq b$ sont triviales. $\square$

La proposition ci-dessus nous assure que si une clause exclut l'intervalle $|0, 1/8|$ et qu'une autre exclut $|1/8, 1/4|$, nous pouvons dire que la solution ne se trouve pas sur l'intervalle $[0, 1/4]$. De proche en proche, il semble donc possible de restreindre itérativement l'espace des solutions admissibles. Il faut toutefois parvenir à déterminer quels sont les intervalles exclus par une clause $\neg(\neg v_i \wedge \neg v_j \wedge \neg v_k)$ quelconque. Idéalement, nous voudrions décrire ces intervalles à l'aide d'une expression analytique facile à manipuler.

Définition 25. Soit une clause $\neg(\neg l_i \wedge \neg l_j \wedge \neg l_k)$ du problème 3-SAT_{ω_0} , avec $i < j < k$. Posons $\rho_i, \rho_j, \rho_k \in \{0, 1\}$ tels que, pour tout $m \in \{i, j, k\}$, $\rho_m = 0$ si le littéral l_m est un littéral positif et 1 sinon. On représente cette clause à l'aide d'un sextuplet $C = (i, j, k, \rho_i, \rho_j, \rho_k)$. Lorsque le signe des littéraux est sans importance, on écrira plus simplement $C = (i, j, k)$.

Proposition 14. Une clause $C = (i, j, k, \rho_i, \rho_j, \rho_k)$ exclut un ensemble de 2^{k-3} intervalles de mesure $1/2^k$:

$$\bigcap_{m \in \{i, j, k\}} \left(\bigcup_{n=1}^{2^{m-1}} \left| \frac{2(n-1) + \rho_m}{2^m}, \frac{2(n-1) + \rho_m + 1}{2^m} \right| \right)$$

Démonstration. Les intervalles d'exclusion de C sont les points qui rendent les littéraux l_i, l_j et l_k faux tous les trois. Nous pouvons voir ces intervalles comme l'intersection de l'ensemble des intervalles exclus respectivement par $\neg l_i, \neg l_j$ et $\neg l_k$. Or l'ensemble des intervalles exclus par un littéral $\neg l_m$ quelconque (avec $m \in \{i, j, k\}$) se porte sur toutes les séquences $(v_l)_{l \in \mathbb{N}^*}$ telles que $\nu(v_m) = \rho_m$. Nous voulons donc exclure toutes les séquences de $\{0, 1\}^{\mathbb{N}^*}$ dont la représentations géométrique est :

$$\phi((v_l)_{l \in \mathbb{N}^*}) = \sum_{l=1}^{m-1} \frac{\nu(v_l)}{2^l} + \frac{\rho_m}{2^m} + \sum_{l=m+1}^{\infty} \frac{\nu(v_l)}{2^l}.$$

On remarque que, quelles que soient les valeurs de $\nu(v_l)$ pour $l > m$, la somme de droite est bornée par

$$0 < \sum_{l=m+1}^{\infty} \frac{\nu(v_l)}{2^l} < \sum_{l=m+1}^{\infty} \frac{1}{2^l} = \frac{1}{2^m}$$

Les séquences que nous voulons exclure se trouvent donc sur les intervalles suivants :

$$\begin{aligned} & \bigcup_{(v_i)_{i \in \mathbb{N}^*}} \left| \sum_{i=1}^{m-1} \frac{\nu(v_i)}{2^i} + \frac{\rho_m}{2^m} + 0, \sum_{i=1}^{m-1} \frac{\nu(v_i)}{2^i} + \frac{\rho_m}{2^m} + \frac{1}{2^m} \right| \\ &= \bigcup_{n=1}^{2^{m-1}} \left| \frac{2(n-1) + \rho_m}{2^m}, \frac{2(n-1) + \rho_m + 1}{2^m} \right| \end{aligned}$$

On conclut en prenant l'intersection de ces intervalles pour $m = i, m = j$ et $m = k$. $\square$

Bien que l'expression que nous venons de dériver ne soit pas particulièrement aisée à manipuler, elle présente une symétrie qu'il est important de relever. Tout d'abord, on observe que la largeur et le nombre d'intervalles ne dépend que de l'indice k correspondant au plus grand littéral l_k , si bien que les clauses $\neg(l_3 \wedge \neg l_4 \wedge l_{11})$ et $\neg(\neg l_6 \wedge l_8 \wedge \neg l_{11})$, par exemple, compteront tous deux 2^8 intervalles de longueur 2^{-11} . Aucun de ces intervalles ne se touchera, mais l'écart entre deux intervalles consécutifs ne pourra prendre que trois valeurs possibles : $2/2^i, 2/2^j$ ou $2/2^k$. Si k est ainsi responsable du nombre et de la taille des intervalles, les indices i et j , quant à eux, définissent leurs distributions le long du segment de droite $[0, 1]$. Quant aux signes des littéraux, capturés par ρ_i, ρ_j et ρ_k , ils ont pour effet de décaler les intervalles d'une distance $2^{-i}, 2^{-j}$ et 2^{-k} (respectivement). On remarque en particulier qu'une clause présentant au moins un littéral positif ne pourra jamais exclure le segment $[0, 1/2^k]$. Réciproquement, une clause présentant au moins un littéral négatif ne pourra pas exclure le segment $[1 - 1/2^k, 1]$. Comme il existe huit façons de choisir ρ_i, ρ_j et ρ_k , il faudra toujours un minimum de huit clauses de même indice k pour exclure l'ensemble du segment $[0, 1]$.

Outre ces propriétés très générales, il est curieux de constater que les intervalles d'exclusion obéissent à une évolution fractale. Pour s'en rendre compte, il peut être utile de représenter les clauses $C_0 = (i, j, k, \rho_1, \rho_2, \rho_3)$ et $C_1 = (i+1, j+1, k+1, \rho_1, \rho_2, \rho_3)$ sur un graphique. On utilisera la fonction indicatrice $\mathbb{1}_{C_1}$ qui prend la valeur 1 pour tous les réels appartenant à un intervalle d'exclusion, et la valeur 0 dans le cas contraire. On identifiera alors rapidement une symétrie intéressante. Le graphique obtenu pour C_1 sur le segment $[0, 1/2]$ est identique à celui de $C_0 = (i, j, k, \rho_1, \rho_2, \rho_3)$ sur l'ensemble du segment $[0, 1]$, à un facteur d'échelle près. En fait, tout se passe comme si C_1 se dupliquait en 2 exemplaires de C_0 , qui se répartissaient de part et d'autre du point $1/2$. Si l'on continue l'expérience avec $C_3 = (i+3, j+3, k+3, \rho_1, \rho_2, \rho_3)$, on remarquera exactement le même phénomène. C_3 présente en effet quatre duplications de C_0 , réparties symétriquement sur les quatre quarts de segment de $[0, 1]$. En généralisant, on peut dire que la clause C_0 suit le même pattern sur le segment $[0, 1]$ que la clause

$C_n = (i+n, j+n, k+n, \rho_1, \rho_2, \rho_3)$ sur les segments $[m/2^n, (m+1)/2^n]$ (avec $0 < m < 2^n$)

Tirant profit de cette périodicité, on pourrait se demander s'il existe un moyen de représenter les intervalles d'exclusion à l'aide d'une formule compacte ? L'expression analytique de la proposition [14](#) est en effet pénible à l'usage, à cause de ses très nombreux termes. La fonction indicatrice $\mathbb{1}_C$, elle non plus, ne résiste pas longtemps à l'épreuve du calcul. Plutôt que de vérifier chaque fois si un $x \in [0, 1]$ appartient, ou non, à l'un des 2^{k-3} intervalles d'exclusion de la clause $C = (i, j, k)$, il serait plus commode d'introduire une fonction périodique.

Définition 26. Soient trois entiers positifs $i < j < k$ et $\rho_i, \rho_j, \rho_k \in \{0, 1\}$. L'harmonique $\psi = (i, j, k, \rho_i, \rho_j, \rho_k)$ est la fonction de $[0, 1]$ dans $\mathbb{R}$:

$$\psi_{i,j,k}(x) = \min_{m \in \{i,j,k\}} \{\sin(2\pi 2^m x + \rho_m \pi)\},$$

Quand elles existent, on note ψ^- et ψ^+ les dérivées de ψ à gauche et à droite. Le nombre $2/2^k$ est appelé la longueur d'onde de ψ et noté λ .

L'harmonique Ψ est une fonction sinusoïdale par parties. Le comportement de ses parties négatives est assez chaotique, mais ses parties positives dessinent exactement les intervalles d'exclusion associés à la clause $C = (i, j, k, \rho_i, \rho_j, \rho_k)$. Une analyse attentive permet d'identifier trois périodes $\lambda_i = 2/2^i$, $\lambda_j = 2/2^j$ et $\lambda_k = 2/2^k$, qui sont responsables de la dispersion et des symétries de l'harmonique. La "longueur d'onde" d'intérêt est λ_k , car c'est elle qui définit l'écart minimum entre chaque intervalle d'exclusion, mais aussi la longueur de ces intervalles. Pour cette raison, nous parlerons souvent de longueur d'onde λ , sans mention explicite de l'indice k , chaque fois que le contexte est suffisamment clair.

Proposition 15. Soit une clause $C = (i, j, k, \rho_i, \rho_j, \rho_k)$. Notons $|a_n, b_n|$ le n -ième intervalle d'exclusion de C , avec $a_1 < a_2 < \dots < a_{2^{k-3}} \in \mathbb{R}$ et $b_1 < b_2 < \dots < b_{2^{k-3}} \in \mathbb{R}$. Pour tout $x \in [0, 1]$, l'harmonique $\psi = (i, j, k, \rho_i, \rho_j, \rho_k)$ vérifie :

- $\psi(x) > 0$ ssi il existe un n tel que $x \in (a_n, b_n)$;
- $\psi(x) = 0$ et $\psi^+ > 0$ ssi il existe un n tel que $x = a_n$;
- $\psi(x) = 0$ et $\psi^- < 0$ ssi il existe un n tel que $x = b_n$.

Dans les deux derniers cas de figure, on dira que x est un point limite gauche (respectivement droit) de ψ .

Démonstration. Choisissons $m \in \{i, j, k\}$. D'après la proposition [14](#), les intervalles d'exclusion du littéral $\neg l_m$ sont donnés par :

$$\bigcup_{n=1}^{2^{m-1}} \left| \frac{2(n-1) + \rho_m}{2^m}, \frac{2(n-1) + \rho_m + 1}{2^m} \right|$$

Il est facile de vérifier que $\sin(2\pi 2^m x + \rho_m \pi) > 0$ si et seulement s'il existe un $n \geq 1$ tel que $2(n-1)\pi < 2\pi 2^m x + \rho_m \pi < 2n\pi$.

□

Cette proposition démontre qu'à toute clause C correspond une unique harmonique ψ . De même, à tout intervalle d'exclusion $|a, b|$ rejetant une partie de l'espace de Cantor $\{0, 1\}^{\mathbb{N}^*}$, correspond un intervalle (a, b) rejetant un sous-ensemble de la droite réelle. Par abus de langage, on écrira indifféremment $|a, b|$ ou (a, b) pour désigner un intervalle d'exclusion, chaque fois que le contexte est suffisamment clair.

11.4 Méthode de résolution

Pour résoudre 3-SAT_{N₀}, nous allons procéder à un balayage intelligent du segment $[0, 1]$, en partant d'un itéré initial $x_0 = 0$ et en se déplaçant vers la droite chaque fois qu'une harmonique ψ rejetant cet itéré est rencontrée. Si, en parcourant les clauses du problème une à une, aucune d'entre elles ne rejette x_0 , alors ce dernier est une solution du problème et la méthode s'achève. Dans le cas contraire, cela veut dire que x_0 est un point limite gauche d'une sinusoïde ψ de période λ . Nous pouvons donc prendre $x_1 = x_0 + \lambda/2$ et recommencer la procédure.

Si plusieurs sinusoïdes $\psi_{i,j,k}$ rejettent x_0 , il faut prendre celle dont la période est la plus grande ou — ce qui revient au même — celle dont l'indice k est le plus faible. En procédant de cette manière, on fera évoluer l'algorithme plus rapidement. De plus, en notant $\psi_0, \psi_1, \dots$ les sinusoïdes d'indice $k_1, k_2, \dots$ rejetant successivement les itérés $x_0, x_1, \dots$, on s'assurera que, à toute étape $\tau \subset \omega_1$ de l'algorithme, l'itéré $x_\tau = x_{\tau+1} + \lambda_\tau/2$ est toujours un point limite gauche de $\psi_\tau(x_\tau)$ et un point limite droite de $\psi_{\tau+1}(x_\tau) = 0$. Autrement dit, ψ_τ et $\psi_{\tau+1}$ se croisent en x_τ . Si cette propriété n'est pas claire au lecteur pour l'instant, elle fera l'objet d'une démonstration rigoureuse. Pour le moment, contentons-nous de remarquer que, sous réserve que cette propriété soit bien exacte, les intervalles d'exclusion $|x_0, x_1|, |x_1, x_2|, \dots, |x_{\tau-1}, x_\tau|$ parcourus par notre méthode après τ itérations pourront s'écrire plus simplement $[0, x_\tau]$.

— Algorithme pour $3\text{-SAT}_{\aleph_0}$ —

Soit Φ , l'ensemble des harmoniques associées aux clauses du problème et $x = 0$ l'itéré initial de la méthode.

```

Algorithme ( $\Phi, x$ )
   $\lambda\text{-max} := 0$ 
  foreach ( $\psi, \lambda$ ) in  $\Phi$ 
    if ( $\psi(x)$  rejette  $x$ )
       $\lambda\text{-max} := \text{Max}(\lambda, \lambda\text{-max})$ 
  if  $\lambda\text{-max} > 0$ 
     $x := x + \lambda/2$ 
    if  $x = 1$ 
      return "no solution"
    else
      return Algorithme( $\Phi, x$ )
  return  $x$ 
end

```

Par souci d'efficacité, on pourrait trier les harmoniques ψ par ordre décroissant de longueur d'onde, en lançant une instruction du type *trier*(Φ) au début de l'algorithme. De cette façon, on aurait la garantie que la première harmonique ψ rencontrée qui exclut l'itéré en cours est effectivement celle de plus grande période. Cela nous permettrait de sortir immédiatement de la boucle *foreach*. Notons que le classement par ordre décroissant, en tant que tel, requière, au plus, $\aleph_0$ étapes. En effet, il consiste à vérifier toutes les paires de contraintes. Or les paires de contraintes peuvent être mises en bijections avec $\mathbb{N}^2$ et, en utilisant la fonction couplage de Cantor, il est possible de mettre $\mathbb{N}^2$ en bijection avec $\mathbb{N}$.

Avant de montrer la convergence de la méthode, remarquons que celle-ci pourrait s'arrêter sur un itéré x_τ brisant l'injectivité de ϕ . Dans ce cas, la question doit être adressée de savoir à quelle séquence $v \in \{0, 1\}^{\mathbb{N}^*}$ correspond cet itéré ?

Définition 27. On dit d'un réel $x \in [0, 1]$ qu'il est une solution de $3\text{-SAT}_{\aleph_0}$ à gauche (respectivement à droite) s'il n'est le point limite droit (respectivement gauche) d'aucune harmonique. Une solution est dite à gauche et à droite si elle n'est le point limite d'aucune harmonique.

Lorsque x est une solution à gauche, mais pas à droite, cela veut dire qu'il est l'image par ϕ d'une séquence $v \in \{0, 1\}^{\mathbb{N}^*}$ se terminant par une infinité de 1. Cette séquence est sur le bord gauche d'un intervalle du type $|\phi(v), b|$ mais pas sur le bord droit d'un intervalle du type $|a, \phi(v)|$, ni à l'intérieur d'un quelconque intervalle. Inversement, si x est solution à droite, mais pas à gauche, la séquence $v \in \{0, 1\}^{\mathbb{N}^*}$ dont il est l'image se termine par une

infinité de 0. Elle est sur le bord d'un intervalle d'exclusion de la forme $|a, \phi(v)|$.

Proposition 16. *Soient deux contraintes C_{i_1, j_1, k_1} et C_{i_2, j_2, k_2} avec $0 < k_1 \leq k_2$. Notons respectivement $(I_n)_{n \leq k_1-2}$ et $(J_n)_{n \leq k_2-2}$ les collections de sous-intervalles d'exclusion disjoints associées à ces contraintes. S'il existe $l \leq k_1 - 2$ et $m \leq k_2 - 2$ tels que $I_l \cap J_m \neq \emptyset$, alors $J_m \subseteq I_l$.*

Démonstration. Par la proposition [14](#), les sous-intervalles I_l et J_m s'écrivent

$$I_l = \left| \frac{(n-1)}{2^{k_1}}, \frac{n}{2^{k_1}} \right| \quad J_m = \left| \frac{(n'-1)}{2^{k_2}}, \frac{n'}{2^{k_2}} \right|$$

pour des entiers positifs n et n' non nuls. Supposons que $I_l \cap J_m \neq \emptyset$ mais que J_m ne soit pas un sous-ensemble de I_l . Il faut soit que $(n'-1)/2^{k_2} < n/2^{k_1} < n'/2^{k_2}$, soit que $(n'-1)/2^{k_2} < (n-1)/2^{k_1} < n'/2^{k_2}$. Or comme $k_1 \leq k_2$, il existe un entier positif p tel que $2^{k_1} = 2^{k_2}/p$. En injectant dans les relations plus haut et en simplifiant, il faut soit que $(n'-1) < np < n'$, soit que $(n'-1) < (n-1)p < n'$. Aucune de ces deux relations n'est satisfaisable, car np et $(n-1)p$ sont des entiers positifs. On en conclut que $J_m \subseteq I_l$. $\square$

Proposition 17. *Soit $x_\tau \in [0, 1]$, le point obtenu après $\tau \subset \omega_1$ itérations de l'algorithme. Supposons que la solution ne soit pas dans $[0, x_\tau]$ et que toutes les harmoniques ψ qui rejettent x_τ à gauche ont ce point pour limite gauche. Dans ce cas, exactement une des deux propositions ci-dessous est correcte :*

1. x_τ est solution du problème à gauche ;
2. À la prochaine étape de l'algorithme, la solution du problème n'est pas dans $[0, x_{\tau+1}]$ et toutes les harmoniques d'exclusion ψ qui rejettent $x_{\tau+1}$ à gauche ont ce point pour limite gauche.

Démonstration. Considérons la famille $(\psi)_i$ des harmoniques d'exclusion rejetant x_τ à gauche. Si cette famille est nulle, le point 1 est vrai. Sinon, on note ψ_τ celle de ces harmoniques avec la plus grande longueur d'onde λ_τ . Par hypothèse, toute harmonique $\psi_j \in (\psi)_i$ a x_τ comme point limite gauche et admet donc le sous-intervalle d'exclusion $I_j = (x_\tau, x_\tau + \lambda_j/2)$ (où λ_j est la longueur d'onde de ψ_j). Par continuité de ψ_j et ψ_τ , il existe donc un réel $\epsilon_j > 0$ tel que $x_\tau + \epsilon \in (I_j \cap I_{x_\tau})$. Par la proposition précédente, on a donc $I_j \subseteq I_\tau$. Cela démontre que la solution n'est pas dans $[x_\tau, x_\tau + \lambda_\tau/2]$. Or par hypothèse, on sait aussi que la solution n'est pas dans $[0, x_\tau]$. Le prochain itéré $x_{\tau+1} = x_\tau + \lambda_\tau/2$ donné par l'algorithme est donc tel que $[0, x_\tau] \cup [x_\tau, x_\tau + \lambda_\tau/2] = [0, x_{\tau+1}]$ (proposition [13](#)). Cela prouve la prémisse du point 2.

Supposons maintenant qu'il existe une harmonique ψ_j rejetant $x_{\tau+1}$ à gauche et n'ayant pas $x_{\tau+1}$ comme point limite gauche (il rejette donc aussi

$x_{\tau+1}$ à droite). Notons I_j le sous-intervalle d'exclusion de ψ_j contenant $x_{\tau+1}$. Par continuité de ψ_j et ψ_τ , il existe un réel $\epsilon_j > 0$ tel que $x_{\tau+1} - \epsilon \in (I_j \cap I_{x_\tau})$. Par la proposition précédente, il faut donc soit que $I_j \subseteq I_\tau$ (si $\lambda_j < \lambda_\tau$), soit que $I_\tau \subseteq I_j$ (si $\lambda_j > \lambda_\tau$). Dans le premier cas, cela contredit le fait que ψ_j rejette $x_{\tau+1}$ à gauche, car $x_{\tau+1}$ est un point limite à droite de $I_\tau = (x_\tau, x_\tau + \lambda_\tau/2)$. Dans le deuxième cas, c'est incompatible avec la construction de ψ_τ . Sans perte de généralité en effet, soit I_j n'a pas x_τ comme point limite à gauche, soit il l'a. S'il ne l'a pas, alors $I_\tau \subseteq I_j$ ne l'a pas non plus et cela contredit le fait que ψ_τ appartienne à la famille $(\psi)_i$ définie au début de la preuve. Mais si I_j a x_τ comme point limite à gauche, alors ψ_j appartient aussi à $(\psi)_i$ et λ_τ n'est plus l'harmonique de plus grande longueur d'onde de cette famille (puisque $\lambda_\tau < \lambda_j$). $\square$

Géométriquement, la proposition ci-dessus revient à dire que les fonctions $\psi_0, \psi_1, \dots, \psi_\tau$ construites par la méthode ont les itérés $x_0, x_1, \dots, x_\tau$ comme limites à gauches. De proche en proche, ces fonctions construisent un intervalle d'exclusion $[x_0, x_1] \cap [x_1, x_2] \cap [x_{\tau-1}, x_\tau] = [x_0, x_\tau]$ de plus en plus grand, jusqu'à atteindre 1 ou à trouver une solution. Il reste cependant à se demander ce qu'il advient lorsque les longueurs d'onde $\lambda_1, \lambda_2, \dots$ des fonctions $\psi_1, \psi_2, \dots$ rencontrées par la méthode à chaque itération tendent vers 0? Le point x obtenu au bout de cette limite est-il une solution? Dans le cas contraire, nous n'avons pas lieu de penser, *a priori*, que cette situation ne puisse pas se produire un grand nombre de fois, au point, peut-être, d'empêcher la méthode de converger en un temps dénombrable.

Proposition 18. *Après $\tau \subset \omega_1$ itérations, supposons que l'algorithme produise une suite infinie d'itérés $(x_{\tau+n})_{n \in \mathbb{N}}$ tels que $\lim_{n \rightarrow \infty} x_{\tau+n} = x$. Si la solution n'est pas dans $[0, x_\tau]$ et que toutes les harmoniques d'exclusions ψ qui rejettent x_τ à gauche ont ce point pour limite gauche, alors x est solution du problème à droite.*

Démonstration. Pour tout $n \in \mathbb{N}$, considérons les deux propriétés suivantes : (i) la solution n'est pas dans $[0, x_{\tau+n}]$ et (ii) toutes les harmoniques d'exclusions ψ qui rejettent $x_{\tau+n}$ à gauche ont ce point pour limite gauche. Par hypothèses, (i) et (ii) sont vraies pour $n = 0$. La proposition [17](#) nous dit que si ces propriétés sont vraies pour $x_{\tau+n}$, alors elle est vraie pour $x_{\tau+n+1}$. Par récurrence, on conclut que, pour tout $n \in \mathbb{N}$, tout $x_{\tau+n}$ possède les propriétés (i) et (ii).

Supposons maintenant par l'absurde qu'il existe une harmonique ψ rejetant x à droite. Par continuité de ψ , il existe donc un $\epsilon > 0$ tel que tout réel $y \in (x - \epsilon, x)$ soit rejeté par ψ à droite et à gauche. Comme $\lim_{n \rightarrow \infty} x_{\tau+n} = x$, il existe un $N \in \mathbb{N}$ tel que $x_{\tau+N} \in (x - \epsilon, x)$. Ce point $x_{\tau+N}$ est donc rejeté par ψ à gauche et à droite. Autrement dit, ψ rejette ce point à gauche sans l'avoir pour limite à gauche, ce qui contredit la propriété (ii). On conclut que x est bien une solution du problème à droite.

□

Nous sommes maintenant en mesure de prouver itérativement la validité de l'algorithme. À la première étape, l'itéré $x_0 \in [0, 1]$ est nul. Si aucune contrainte ne le rejette, c'est donc une solution au problème. Dans le cas contraire, l'hypothèse de la proposition 17 est vérifiée, puisque 0 est un point limite à gauche de tous les intervalles rejetant 0. Par récurrence, et en utilisant la proposition 17, on conclut que, à chaque étape τ de l'algorithme, soit x_τ est une solution du problème, soit il se trouve sur le bord gauche d'une harmonique d'exclusion ψ_τ de longueur d'onde λ_τ . Dans ce dernier cas, on peut alors relancer l'algorithme en prenant $x_{\tau+1} = x_\tau + \lambda_\tau/2$. Dans le cas particulier où $x_{\tau+1} = 1$, la proposition 17 nous enseigne que la solution n'est pas dans $[0, x_\tau + \lambda_\tau] = [0, 1] = [0, 1]$ (la dernière égalité vient de la proposition 13). Il est donc facile de voir que x_τ va croître à chaque étape de l'algorithme, jusqu'à atteindre 1 ou à s'arrêter sur une solution.

Il reste encore à montrer que la convergence se produit, au plus, après $\mathcal{O}(\omega_0)$ itérations. Pour ce faire, commençons par remarquer que, à chaque appel de la boucle *while* de l'algorithme, ce dernier effectue, au plus, un nombre d'opérations proportionnel au nombre de clauses, qui est lui-même infini dénombrable. De plus, on peut mettre en bijection chaque appel de la boucle *while* avec l'itéré x_τ calculé lors de cet appel. Or tant que x_τ est rejeté par une clause, x_τ est un nombre rationnel. En effet, on commence avec $x_0 = 0$ (un rationnel) et on ajoute, à chaque itération, le nombre $\lambda_\tau/2 = 1/2^k$ (pour un certain $k \in \mathbb{N}$), qui est aussi un rationnel. De plus, il n'existe aucun $\tau' < \tau$ tel que $x_\tau = x_{\tau'}$. Si cela était le cas, on violerait la règle de construction de l'algorithme, qui pose $x_{\tau+1} = x_\tau + \lambda_\tau/2 > x_\tau$. Par conséquent, le τ -ième appel de la boucle *while*, s'il ne converge pas vers une solution, peut être associé à un nombre rationnel x_τ distinct. En d'autres termes, il existe une injection entre l'ensemble des étapes de l'algorithme et l'ensemble des nombres rationnels. Comme $\mathbb{Q}$ est de cardinalité $\aleph_0$, on conclut que l'algorithme est de complexité, au plus, $\mathcal{O}(\omega_0)$. Ce raisonnement est résumé dans le théorème suivant.

— Solubilité de 3-SAT $_{\aleph_0}$ —

Le problème 3-SAT $_{\aleph_0}$ est soluble dans le pire des cas en un nombre, au plus, infini dénombrable. Dans le cas particulier où le problème ne comporte aucune solution, il est soluble en temps fini.

Remarquons que si les clauses ont été triées par ordre décroissant de longueur d'onde, un problème 3-SAT $_{\aleph_0}$ n'admettant aucune solution pourra s'achever en un nombre fini d'étapes. En effet, chaque appel de la boucle *while* prendra un temps fini. De plus, il ne peut y avoir qu'un nombre fini

d'appel de cette boucle, sans quoi cela voudrait dire que les itérés $v_0, v_1, \dots$ formeraient une série infinie. Or la proposition 18 démontre qu'une telle série converge toujours vers une solution. Si maintenant les clauses ne sont pas présentées par ordre croissant de longueurs d'ondes, le programme pourrait toujours décider de garder par défaut la première harmonique ψ rejetant l'itéré x_τ qu'il rencontrerait. Notons (a, b) l'intervalle d'exclusion de ψ dans lequel (ou sur le bord duquel) se trouve x_τ à la τ -ième itération. Au lieu de faire $x_{\tau+1} = x_\tau + \lambda$, l'algorithme poserait $x_\tau = b$. Dans ces conditions, il convergerait également vers une solution, quoique moins "efficacement" (la notion d'efficacité est toutefois relative dans le contexte de machines transfinies, en raison de l'équipotence des ordinaux dénombrables). La preuve exige toutefois que les clauses soient d'ordinalité ω_0 . Si ce n'est pas le cas, nous savons, par la proposition 9, qu'il est possible de simuler une machine comportant des entrées de taille supérieures à ω_0 en se limitant à un ruban de taille ω_0 . Cela exige simplement de présenter le problème $3\text{-SAT}_{\aleph_0}$ sous une forme adéquate, sans constituer une restriction d'ordre proprement conceptuel.

Proposition 19. *L'impossibilité du problème $3\text{-SAT}_{\aleph_0}$ possédant exactement ω_0 clauses et ω_0 variables peut être déterminée en un temps fini.*

Démonstration. La preuve découle de ce que l'on vient de dire. $\square$

On en conclut qu'un programme fini peut déterminer, en un temps fini, si le problème $3\text{-SAT}_{\aleph_0}$ admet une solution. A fortiori, cela implique aussi qu'il peut dire si une machine transfinie exécutant ce programme s'arrête en temps fini. Nous avons ainsi exhibé l'exemple d'un programme fini capable de déterminer si une machine de Turing infinie $\mathbf{T}_1$ s'arrête. Cette machine serait la restriction de $\mathbf{T}_1$ à un nombre fini de clauses et de variables propositionnelles, dont le comportement serait identique à celui de $\mathbf{T}_1$ en un temps fini. En d'autres termes, cela veut dire qu'un problème $3\text{-SAT}_{\aleph_0}$ impossible ne contient pas plus d'information qu'un certain problème 3-SAT judicieusement choisi. Ou encore, qu'il existe un nombre fini de clauses qui, prises ensembles, excluent toutes les solutions admissibles. Les clauses et les variables supplémentaires ne représentent, dans le meilleur des cas, qu'une redondance inutile.

Une question que l'on pourrait se poser est celle de savoir si un problème $3\text{-SAT}_{\aleph_0}$ impossible, dont on a extrait le sous problème 3-SAT impossible, deviendrait possible ou non ? Cela exige de définir exactement ce qu'on entend par "extraction". Sans doute s'agirait-il d'enlever le sous-ensemble de clauses mutuellement exclusives parcourues par la machine finie pour déterminer l'impossibilité de $3\text{-SAT}_{\aleph_0}$, ainsi que toutes les variables propositionnelles qui, à la suite à cette extraction, demeureraient libres de prendre n'importe quelles valeurs. À procéder de cette façon itérativement, arriverait-on

sur un problème $3\text{-SAT}_{\aleph_0}$ soluble ? Si le nombre d'extractions est fini, cela va sans dire, car l'union finie d'ensembles finis de clauses donne un nombre fini de clauses. Or $3\text{-SAT}_{\aleph_0}$ est composé d'un nombre de clauses infini dénombrable. En revanche, il se pourrait qu'en procédant à une infinité d'extractions, on finisse par enlever l'ensemble des clauses ou à réduire $3\text{-SAT}_{\aleph_0}$ à un problème 3-SAT fini et soluble. Ce nombre infini d'extractions n'est pas réalisable par un programme fini, mais bien par une machine $\mathbf{T}_1$. Dès lors, on pourrait écrire un programme infini qui détermine si un problème du type $3\text{-SAT}_{\aleph_0}$ est soluble ou non en un temps infini. Cette ligne de réflexion représente une preuve alternative, et une confirmation, de la solubilité de $3\text{-SAT}_{\aleph_0}$ par un programme transfini. Ce que nous souhaiterions faire à présent, c'est montrer que le concept de machine transfinie $\mathbf{T}_1$ que nous avons introduite au début du document peut implémenter ce type de programmes infinis dont nous n'avons, jusqu'à présent, donné que le pseudocode.

Proposition 20. *Le problème $3\text{-SAT}_{\aleph_0}$ est soluble par une machine de Turing transfinie $\mathbf{T}_1 = (K, \Sigma, \delta, s, H, \alpha)$.*

Démonstration. Nous savons qu'il existe un encodage permettant d'exprimer les clauses et les variables du problème à l'aide d'une chaîne de caractères $x \in \Sigma^*$, car c'est le cas pour une machine de Turing classique. Nous pouvons supposer, sans perte de généralité, que x est de taille égale à ω_0 . La proposition [9] stipule en effet qu'une machine de Turing utilisant un ruban de taille supérieure peut être simulée par une machine utilisant un ruban de longueur ω_0 . Dans ces conditions, la proposition [19] nous enseigne que, si le programme n'admet pas de solutions, il existe une méthode qui détermine cet état de fait en un temps fini $\tau \subset \omega_1$. Par la proposition [6], une machine transfinie $\mathbf{T}_1$ finissant en un temps fini se comporte comme sa restriction finie. Or implémenter l'algorithme [11.4] sur une machine finie ne fait de difficulté. Nous pouvons déduire que notre machine $\mathbf{T}_1$ n'atteindra $|_0\rangle$ que si le problème $3\text{-SAT}_{\aleph_0}$ admet une solution. Il suffit donc de donner pour instruction que la machine s'arrête dès qu'elle atteint le symbole $|_0\rangle$, que ce soit dans l'état ∞ ou dans n'importe quel autre. En vertu de la proposition [18] l'encodage de l'itéré x_{ω_0} , présent sur le ruban à ce moment-là, est la solution du problème. $\square$

11.5 Programme pour une machine transfinie

Bien que les harmoniques ψ soient élégantes pour comprendre la méthode et en prouver la convergence, elles ne sont pas des plus commodes en termes de calculs informatiques. Il est plus difficile de calculer une sinusoïde que de vérifier une contrainte binaire de la forme $l_i \vee l_j \vee l_k$. De plus, le calcul du sinus peut ne pas être exact s'il est tronqué avant la ω_0 décimale, si bien qu'il faudrait jusqu'à ω_0 étapes pour vérifier une clause de $3\text{-SAT}_{\aleph_0}$, alors que cela pourrait être fait en trois étapes.

Pour éviter de manipuler des sinusoïdes, le pseudocode va construire, à chaque étape $\tau \subset \omega_1$, un vecteur v_τ à valeurs dans $\{0, 1\}^{\aleph}$, qui correspond à une solution de $3\text{-SAT}_{\aleph_0}$. Le programme vérifie une clause $l_i \vee l_j \vee l_k$ en posant $\nu(v_i)$, $\nu(v_j)$ et $\nu(v_k)$ égaux aux i -ème, j -ème et k -ème coordonnée du vecteur v_τ en cours. Les clauses, quant à elles, sont représentées dans une matrice C de dimension $\aleph_0 \times 6$, qui est donnée en entrée à l'algorithme. Chaque ligne de cette matrice s'écrit sur la forme $(n_1, n_2, n_3, a_1, a_2, a_3)$. Les entiers $n_1, n_2, n_3 \in \mathbb{N}^*$ réfèrent aux indices de trois variables propositionnelles $v_{n_1}, v_{n_2}, v_{n_3}$ et les nombres binaires $a_1, a_2, a_3 \in \{0, 1\}^{\aleph^*}$ dénotent le sens des littéraux (positifs ou négatifs), selon la même convention que celle utilisée dans la proposition [15](#). La ligne $(2, 4, 6, 0, 1, 0)$, par exemple, représente la contrainte $\neg(\neg v_2 \wedge v_4 \wedge \neg v_6)$ ².

Chaque fois que le programme rencontre une contrainte d'indice k excluant la solution, il va procéder à une addition binaire $v_{\tau+1} = v_\tau + v_k$, où $v_k \in \{0, 1\}^{\aleph^*}$ désigne un vecteur dont toutes les coordonnées sont nulles, à l'exception de la k -ième. Par exemple, si $v_\tau = 01011000\dots$ et $v_k = 00001000\dots$, alors l'addition binaire donne $01100000\dots$. De cette manière, on a bien $\phi(v_{\tau+1}) = \phi(v_\tau) + 1/k$. Dans le cas particulier où la i -ième coordonnées du vecteur v_τ est égale à 1 pour tout $i \geq k$, l'addition est impossible et l'algorithme prend fin. Géométriquement, cela correspond à une situation où $\phi(v_\tau) + 1/2^k \geq 1$. On peut donc conclure qu'il n'y a pas de solution.

Cela peut se produire, par exemple, si l'algorithme a déjà exclu l'intervalle $[0, 15/16]$ et se demande s'il existe une solution sur le tronçon $[15/16, 1]$. À cet instant, le vecteur v_τ vaut $111100\dots$. Dans la représentation géométrique, ce vecteur correspond en effet à une position sur la droite réelle égale à $1/2 + 1/4 + 1/8 + 1/16 = 15/16$, ce que l'on écrit $\phi(v_\tau) = 15/16$. Sup-

2. Notons, pour la clarté, que la matrice C contient un nombre de lignes égal aux nombre de contraintes du problème 3-SAT généralisé; c'est-à-dire un nombre infini dénombrable. Chaque fois que l'on veut évaluer la contrainte C_i , il faut combiner l'information présente dans la i -ième ligne de la matrice avec les valeurs du vecteur v_τ . Ce vecteur varie au cours de temps. Si une solution existe, le vecteur doit converger vers une solution en un temps, au plus, infini dénombrable. Ceci ne serait pas possible avec un algorithme de force brute, qui essaie toutes les valeurs de τ , mais en lisant les contraintes de façon intelligente, j'ai montré précédemment qu'il était possible de forcer τ à converger en un temps infini dénombrable (si une solution existe) ou à conclure en un temps fini que la solution n'existe pas.

posons que, à l'étape suivante de l'algorithme, on découvre la contrainte $\neg(v_1 \wedge v_2 \wedge v_3)$. Cette contrainte exclut l'intervalle $[7/8, 1]$, si bien qu'il n'y a plus de solution possible.

Du point de vue de l'algorithme, la découverte d'un nouvel intervalle d'exclusion signifie qu'il faut se déplacer sur la droite réelle d'une distance équivalente à la longueur de l'intervalle qui vient d'être exclu, à savoir $1/8$. Techniquement, on doit faire l'opération $v_{\tau+1} = v_{\tau} + v_k$, avec $v_{\tau} = 111100\dots$ et $v_k = 00100000\dots$. L'addition binaire de ces vecteurs revient à faire $\phi(v_{\tau}) + \phi(v_k) = 15/16 + 1/8$, ce qui donne une valeur supérieure à 1. Or il n'existe aucun vecteur v , dans l'espace de Cantor, tel que $\phi(v) > 1$. L'addition est donc impossible et l'algorithme prend fin.

Algorithme (C)

```

v := vectorZero();
trierParOrdreCroissant(C);
search := true
while(search)
  search := false
  for (n1, n2, n3, a1, a2, a3) in C
    if (v(n1) == a1 and v(n2) == a2 and v(n3) == a3)
      search := true
      n := n3
      if(v(n) != 1)
        v(n) := 1
      else if (n > 1)
        n := n-1
      else
        return "pas de solution"
      end
    end
  end
end
return v
end

```

Ce programme corrobore la démonstration d'après laquelle une machine de Turing transfinie $\mathbf{T}_1$ peut résoudre un problème NP-complet généralisé en un temps infini dénombrable $\aleph_0$, et non en un temps infini indénombrable $\aleph_0^2$. Cela veut dire qu'il est tout aussi facile pour $\mathbf{T}_1$ de trouver une solution que de vérifier une solution. Ou pour l'exprimer encore différemment, on peut dire que la classe des problèmes P^T transfinis est égale à la classe des problèmes NP^T , ce que l'on note $P^T = NP^T$.

La première question qui reste en suspens consiste à se demander si les problèmes NP-complets transfinis correspondent à des problèmes réels,

auxquels notre esprit est régulièrement exposé. J'ai en effet montré que les problèmes NP-complets transfinis pouvaient être représentés sur la droite réelle ; que c'étaient des problèmes définis dans le continu. Il n'est toutefois pas sûr que ces problèmes ne soient pas qu'une simple fabrication, un artefact mathématique, conçus spécialement pour les besoins de mon argumentation — et qui n'auraient aucune application pratique dans la vie humaine quotidienne.

La deuxième question d'importance consiste à se demander si notre esprit fonctionne effectivement comme une machine de Turing transfinie. À supposer en effet que les problèmes NP-complets généralisés soient des tâches que l'esprit humain est généralement amené à résoudre, on ne comprend pas pourquoi cette tâche ne pourrait pas être surmontée avec un réseau de neurones, par exemple. L'hypothèse d'après laquelle l'esprit fonctionne comme une machine de Turing est inquiétante pour cette simple et bonne raison que notre cerveau, quant à lui, est de taille finie. C'est bien d'ailleurs pour cette raison que j'ai distingué la notion de modèle de la notion de fonction. Je n'ai pas cessé d'insister sur le fait que la machine de Turing était un modèle de l'intuition, non de l'esprit proprement. Il se peut donc que notre esprit ne fonctionne pas du tout comme une machine de Turing transfinie ; mais il faut néanmoins que son fonctionnement supporte le même type de fonctions que celles dont la machine de Turing transfinie se rend capable.

Je vais envisager ces délicates questions dans le prochain chapitre. Je profiterai également de l'occasion pour dessiner une vision philosophique cohérente du rapport entre la conscience et le cerveau. Mais auparavant, j'aimerais encore toucher un mot à propos de notre résultat central, comme quoi $P^T = NP^T$, et me demander si cela est susceptible de nous enseigner quoi que ce soit de pertinent à propos du problème $P? = NP$. On a l'impression, à première vue, qu'un résultat obtenu sur des problèmes NP-complet transfinis pourrait être transposé, au moins partiellement, au cas des problèmes NP-complets finis. Je vais montrer que rien n'est moins sûr ; je vais toutefois dresser une liste de stratégies prometteuses.

11.6 Stratégies concernant $\mathcal{P} = \mathcal{NP}$

Le mathématicien Charles Fefferman aurait dit un jour : "Écrire une démonstration, c'est jouer aux échecs avec le Diable. Bien que vous perdiez sans cesse, vous avez le droit de revenir en arrière aussi souvent que nécessaire". Vu sous cet angle, il se pourrait qu'étudier un problème mathématique aussi coriace que $\mathcal{P} = \mathcal{NP}$ s'apparente à l'analyse d'une ligne échiquienne bien documentée. Nous savons que la position des pièces ne peut conduire, in fine, qu'à trois issues possibles : la victoire ($\mathcal{P} = \mathcal{NP}$), la défaite ($\mathcal{P} \neq \mathcal{NP}$) ou l'égalité ($\mathcal{P} = \mathcal{NP}$ est indécidable), mais aucun argument décisif ne semble pouvoir être apporté. Heureusement, il existe une différence de taille entre

l'échec et mat et la preuve mathématique pure. Si le premier s'obtient en réduisant la complexité d'une position particulière par des échanges de coups bien pensés (et c'est peut-être ce que cherche à imiter une certaine branche des sciences algorithmiques), la seconde, au contraire, procède par généralisations croissantes (quitte à perdre de vue les objets concrets). Sans doute n'est-il jamais venu à l'esprit d'un joueur d'échecs chevronné de transposer une ouverture difficile sur un échiquier de $\aleph_0 \times \aleph_0$ cases ? Aussi étrange que cela puisse paraître, je pense que, en ce qui concerne la recherche d'une preuve mathématique, formuler un problème NP-complet à l'aide d'un ensemble infini de variables et de contraintes pourrait apporter un éclairage intéressant.

C'est exactement ce que j'ai fait avec le problème 3-SAT. Ce problème est NP-complet pour une machine de Turing classique. Je l'ai toutefois généralisé à un ensemble infini dénombrable de variables propositionnelles $D = \{v_1, v_2, \dots\}$ et de clauses $l_i \vee l_j \vee l_k$ (avec i, j, k des entiers naturels distincts), où le littéral l_i dénote soit la variable v_i (on parle alors de littéral positif) soit $\neg v_i$ (littéral négatif). Toute solution admissible de ce nouveau problème 3-SAT $_{\aleph_0}$ consiste alors à assigner la valeur *true* à un sous-ensemble de variables propositionnelles $V \subseteq D$ et la valeur *false* à son complémentaire V^c . L'espace des solutions possibles, quant à lui, est donné par l'ensemble des parties de $\mathbb{P}(D)$. Dans ce contexte, j'ai construit une machine de Turing transfinie $\mathbf{T}_{\omega_1}$. Cette machine se comporte, à tout égard, comme une machine de Turing classique, si ce n'est qu'elle est capable de réaliser un nombre infini dénombrable d'opérations successives (formellement, on dira que le nombre d'opérations α réalisées doit satisfaire $\alpha < \omega_1$).

Il est facile de voir que notre machine $\mathbf{T}_{\omega_1}$ peut vérifier une solution du problème 3-SAT $_{\aleph_0}$ en un temps infini dénombrable. Plus précisément, on dira que le nombre d'étapes requis peut être mise en bijection avec les $\aleph_0$ clauses définissant le problème. En revanche, pour parcourir l'ensemble des solutions possibles une à une, il faudrait envisager tous les éléments de $\mathbb{P}(D)$, dont la cardinalité est indénombrable. À moins d'avoir "de la chance", ou d'être confronté à un problème admettant un "grand nombre de solutions" (dans un sens que nous préciserons), une machine $\mathbf{T}_{\omega_1}$ ne serait pas capable de résoudre le problème 3-SAT $_{\aleph_0}$ par force brute. Cette situation n'est pas sans rapport avec le problème 3-SAT classique. Il est dès lors naturel de se demander s'il existe un lien quelconque entre les deux questions suivantes :

- Q1 : Existe-t-il une machine de Turing déterministe capable de solutionner n'importe quelle instance de 3-SAT en temps polynomial (dans le pire des cas) ?
- Q2 : Existe-t-il une machine de Turing déterministe *transfinie* capable de solutionner n'importe quelle instance du problème 3-SAT $_{\aleph_0}$ en un nombre infini dénombrable d'opérations (dans le pire des cas) ?

Les développements que j'ai exposés dans les sections précédente conduisent à un éclairage sur cette question. D'une part, ils apportent une réponse positive à la question Q2. D'autre part, et c'est là le point le plus difficile, ils permettent éventuellement de mieux comprendre la relation existant entre les questions Q1 et Q2 (à supposer qu'il en existe seulement une, car ce n'est pas si évident). Plusieurs cas d'espèces sont envisageables. On pourrait penser que si on détenait un algorithme capable de résoudre le problème 3-SAT en temps polynomial sur une machine de Turing classique, alors on pourrait généraliser cet algorithme sur une machine de Turing transfinie de sorte à résoudre le problème 3-SAT $_{\aleph_0}$. Cela voudrait dire que Q1 $\rightarrow$ Q2. En prenant la contraposée, on pourrait alors montrer $\mathcal{P} \neq \mathcal{NP}$ à l'aide d'une simple réfutation de Q2.

Il peut également être suggéré, à rebours de cette première approche, qu'une réponse positive à la question Q2 implique peut-être une réponse positive à la question Q1. Dès lors qu'on est capable de construire un algorithme résolvant le problème 3-SAT $_{\aleph_0}$, il serait intéressant de transposer les principes de fonctionnement de cet algorithme sur une machine de Turing classique, avec l'espoir de résoudre ainsi le problème 3-SAT. Cela aboutirait à une preuve constructive comme quoi $\mathcal{P} = \mathcal{NP}$. Même en l'absence d'une preuve constructive, un simple résultat général montrant que Q2 vraie implique Q1 vraie suffirait à résoudre le problème, pour peu qu'il soit établi, d'une manière ou d'une autre, que Q2 est vraie.

Les situations, intuitivement moins probables, où Q1 vraie implique Q2 faux et Q2 vraie implique Q1 faux, sont tout aussi intéressantes que les deux situations présentées ci-dessus. Il est cependant difficile de concevoir comment l'obtention d'une solution du problème 3-SAT $_{\aleph_0}$ permettrait de conclure à l'impossibilité d'un algorithme polynomial pour le problème 3-SAT classique, à moins de parvenir à exhiber une différence essentielle, entre les deux problèmes, qui expliquent que ce qui fonctionne dans le cas infini est précisément ce qui empêche un algorithme de fonctionner dans le cas fini. Étant donné le comportement, parfois imprévisible, des ensembles infinis, cette voie n'est pas à exclure *a priori* (bien que nous la tenons pour moins fiable).

On peut également étudier ce qu'impliquerait, sur la véracité de Q1, le fait que Q2 soit indécidable — et inversement. Pour cette raison, une question assez naturelle serait de se demander si le problème de l'arrêt est soluble, ou non, pour des machines de Turing transfinie, et sous quelles conditions. Si l'on détenait une machine de Turing transfinie universelle, notée $\mathcal{U}_{\omega_1}$, capable de déterminer si une autre machine de Turing transfinie s'arrête ou non, alors on pourrait peut-être appliquer $\mathcal{U}_{\omega_1}$ à l'entrée finie x correspondant au programme d'une machine Turing finie, et regarder si on peut obtenir la réponse en un temps fini. De cette manière, on aurait résolu le problème de l'arrêt des machines de Turing classiques, ce qui engendrerait

	Stratégies	Q2 vrai	Q2 faux	Q2 indécidable
S1	$Q1 \rightarrow Q2$	-	$\mathcal{P} \neq \mathcal{NP}$	-
S2	$Q2 \rightarrow Q1$	$\mathcal{P} = \mathcal{NP}$	-	-
S3 ?!	$Q1 \rightarrow \neg Q2$	$\mathcal{P} \neq \mathcal{NP}$	-	-
S4 ?!	$\neg Q1 \rightarrow Q2$	-	$\mathcal{P} = \mathcal{NP}$	-
S5	$\not\vdash Q2 \rightarrow \not\vdash Q1$	-	-	indécidable
S6	$\not\vdash Q1 \rightarrow \not\vdash Q2$	décidable	décidable	-
S7 ?!	$\vdash Q2 \rightarrow \not\vdash Q1$	indécidable	indécidable	-
S8 ?!	$\not\vdash Q2 \rightarrow \vdash Q1$	-	-	décidable

TABLE 11.1 – Cette table reprend l’essentiel des stratégies envisageables, à l’exception celles où une indécidabilité / décidabilité impliquerait du vrai / faux (et réciproquement). Les stratégies précédées du symbole " ?!" sont considérées douteuses. Par abus de langage, les caractères $\not\vdash$ et $\vdash$ dénotent respectivement le fait d’être indémontrable (resp. démontrable).

une contradiction. Ce genre de stratégie pourrait s’avérer féconde pour (a) montrer que le problème de l’arrêt existe aussi pour des machines transfinies et (b) tenter de trouver une relation entre la décidabilité de Q1 et Q2.

L’ensemble de ces stratégies est résumé dans le tableau ci-contre. Je m’estime heureux si je suis parvenu, par une analyse attentive du problème, à exclure certaines de ces approches, et à garder ouverte l’une ou l’autre d’entre elles. Notons que la réponse à la question Q2 est susceptible de dépendre, au moins partiellement, de la définition exacte donnée à une machine de Turing transfinie. Dès lors, nous n’excluons pas la possibilité de construire des variantes de machines transfinies (aux comportements éventuellement pathologiques) pour lesquelles Q2 n’aurait pas la même valeur de vérité que celle que j’ai utilisée dans cette thèse. Il se peut que l’une de ces variantes conduisent à des résultats plus pertinents. J’ai toutefois tenté de définir mon concept de la façon la plus générale et la plus opérationnelle qu’il m’était permis de faire. Des recherches plus approfondies méritent d’être menées sur ces importantes questions.

Une des trois fonctions non-calculable de l'intuition humaine consiste à avoir « de la chance » face à un problème NP-complet, de sorte à le résoudre rapidement.

Dans ce chapitre, j'ai montré qu'une machine de Turing transfinie avait systématiquement « de la chance » face aux problèmes NP-complets transfinis. Cela veut dire que, dans le cas transfini, la résolution d'un problème NP-complet n'est pas plus difficile que sa vérification.

J'ai en outre donné une interprétation géométrique du problème NP-complet transfini consistant à trouver les zéros d'une fonction sur la droite réelle. Il en découle qu'une machine de Turing transfinie peut être vue comme une machine capable de penser le continu (en ce compris l'arithmétique et la logique du deuxième ordre).

Il est toutefois important de comprendre que cette machine de Turing transfinie n'est pas un modèle de l'esprit humain, mais un modèle de *l'intuition* humaine (l'intuition du tout et l'intuition de l'infini) ; au même titre que l'ordinateur n'était pas un modèle de l'esprit humain, mais de la *raison* humaine (la pensée pure).

Dans la prochaine partie, j'aimerais me demander ce que la machine de Turing transfinie peut tout de même nous enseigner à propos de la nature de l'esprit humain ? Il reste en effet difficile de comprendre comment un cerveau fini est susceptible d'exécuter des fonctions qui échoient à une machine de Turing transfinie.

Vers une Théorie générale de l'esprit humain

Ce que disent les neurosciences

Chapitre 12

Le Problème Corps-Esprit : Mystérianisme perspectiviste

Au cours de ce petit détour auprès des machines de Turing transfinies, nous nous sommes heurté à un problème de taille. D'une part, j'ai défendu que l'intuition pouvait être conçue de façon explicite sous la forme de trois fonctions non calculables : (1) une fonction de complétion de séries infinies, (2) une fonction de résolution de problèmes indécidables et (3) une fonction « chance » permettant de trouver un bon itéré initial au sein des problèmes NP-complet. D'autre part, j'ai fourni un modèle mathématique supportant conjointement ces trois fonctions : les machines de Turing transfinie.

J'ai toutefois reconnu que la machine de Turing transfinie, bien qu'elle modélise ces trois fonctions de l'intuition, n'est pas pour autant un modèle de l'esprit humain. Il y a en effet une différence de taille entre « modéliser des fonctions de l'intuition » et « modéliser l'esprit humain ». Dans le premier cas, on fournit une machine susceptible de reproduire certains résultats clés de l'intuition, comme le fait d'appréhender l'infini. Dans le second cas, on fournit une machine capable de simuler toutes les dimensions de la pensée, de l'émotionnalité, de la sensation humaine. La différence qui existe entre ces deux machines est du même ordre que la différence séparant irrévocablement une calculatrice (un modèle du raisonnement arithmétique) et un cerveau humain.

Il se peut donc que le modèle de l'esprit humain, s'il existe, n'ait finalement rien de comparable avec les machines de Turing transfinies. La question qui se pose, dès lors, est de comprendre ce que ces machines transfinies peuvent néanmoins nous enseigner de pertinent à propos de la nature de la cognition humaine ? Il y a deux façons de répondre, toute aussi problématiques l'une que l'autre.

1. **Réponse optimiste** : l'esprit humain peut être représenté sous la forme d'un programme implémentable par une machine de Turing

transfinie.

2. **Réponse pessimiste :** l'esprit humain, s'il est modélisable, n'est pas implémentable par une machine de Turing transfinie.

Si l'on opte pour la réponse optimiste, en affirmant que le modèle de l'esprit est un certain programme de la machine de Turing transfinie, c'est-à-dire une succession de lois biophysiques, de propriétés, de relations, modélisables au moyen de la machine de Turing transfinie, on se heurte à une impossibilité physique. De quel droit un système physique de taille fini (notre cerveau, notre organisme...) peut-il effectuer des opérations transfinies ? Ce problème est équivalent à celui qui consisterait à construire une version mécanique concrète de la machine de Turing transfinie, une sorte de superordinateur.

Une manière de contourner ce problème consiste à arguer du fait que la machine du Turing transfinie est certes un modèle de l'esprit humain, mais que c'est un modèle idéalisé. Il est en effet fréquent, en science, de mobiliser des modèles infinis pour représenter des phénomènes finis. Un bon exemple est la théorie euclidienne, qui suppose que les objets de l'espace, tels que des droites, sont infiniment étendues — alors que dans la pratique notre Univers est de taille finie. Le fait que le modèle présuppose l'infini ne semble pas nous gêner outre mesure. Nous pouvons appliquer le modèle infini à des cas finis. L'idéalité du modèle sert d'horizon à la pensée pratique. Elle nous permet de définir et de manipuler plus aisément les objets géométriques

Cette réponse que je fais, qui consiste à voir les machines transfinies comme un modèle idéalisé, n'est toutefois pas tout à fait convaincante. Si les machines de Turing transfinies constituent un modèle idéalisé de l'esprit humain, mais qu'en pratique l'esprit humain ne procède jamais qu'à un nombre fini d'étapes, cela ne reviendrait-il pas à dire que le modèle de l'esprit humain est une machine de Turing classique ? Nous avons vu, en effet, qu'une machine de Turing transfinie qui réalise un nombre d'étapes fini se comporte de façon totalement identique à une machine de Turing classique (voir proposition [6](#)).

C'est pourquoi l'on peut être tenté de dire — et il s'agit de la réponse pessimiste — que l'esprit humain n'est pas modélisé par une machine de Turing transfinie, mais par une machine *finie* ayant un fonctionnement analogue à une machine de Turing transfinie. Il peut sembler contre-intuitif de dire qu'une machine finie puisse se comporter comme une machine transfinie ? Un exemple envisageable serait une machine de Turing classique dotée d'un oracle. Pour rappel, l'oracle est un dispositif, au fonctionnement mystérieux, qui est capable de fournir une réponse à une question prédéfinie. Dans le cas qui nous intéresse, cette question serait par exemple : « quelle est la convergence de cette série infinie ? ». Notre machine se comporterait en tous points comme une machine de Turing classique, mais elle pourrait,

de temps à autre, apporter une réponse à cette question particulière, ce qui la rapprocherait sensiblement du fonctionnement d'une machine de Turing transfinie (j'ai décrit plus précisément le rapport entre la machine de Turing transfinie et les oracles dans la section [10.3.1](#)).

D'après la réponse pessimiste, le modèle de l'esprit serait donc une machine de Turing classique doté d'une espèce de boîte noire susceptible de répondre à des questions sur l'infini. On ne sait pas comment cette boîte fonctionne de l'intérieur ; tout ce que l'on peut dire, c'est (1) qu'elle n'est pas une machine de Turing classique et (2) qu'elle associe, à chaque question, la bonne réponse. Cela revient à dire qu'on ne peut pas expliciter comment l'esprit fonctionne ; qu'il y a en lui un mystère indépassable. Notre réponse pessimiste nous reconduit donc face au dilemme de l'explicitation que j'ai décrit dans le chapitre [9](#). En dernier ressort, cela revient à dire qu'il existe un élément X (l'oracle), disposant de telles et de telles propriétés que l'on ne peut pas expliquer.

Bien sûr, on pourrait tenter de dépasser le mystère de l'élément X en disant que l'oracle, bien qu'inexplicable dans le formalisme d'une machine de Turing classique, puisse être modélisé par un système physique non classique, répondant par exemple aux lois de la mécanique quantique. Il se peut tout à fait que l'étude plus approfondie du cerveau humain nous permette de découvrir un jour, au sein des cellules, des comportements physiques qui ne répondent pas aux lois de la physique classique. S'il est démontré que ces comportements (1) ne sont pas modélisables par une machine de Turing classique et que (2) ils satisfont au moins une des trois fonctions non calculable de l'intuition, on détiendrait une clef de lecture intéressante sur le fonctionnement de la boîte noire de l'intuition.

Cette approche que je viens de décrire, notons-le d'emblée, n'est pas sans évoquer la stratégie de Roger Penrose. Dans son ouvrage *The Emperor's New Mind*, Penrose impute certaines propriétés de la conscience au fonctionnement quantique du cerveau. Il cherche même à décrire un mécanisme le plus précis que possible. Il dépasserait du cadre de cette thèse de discuter les aspects techniques de sa démarche. Il y a tout de même trois raisons très générales pour lesquelles je ne souhaite pas m'aventurer sur les brisées de Penrose. Je vais les exposer ci-dessous.

La première raison pour laquelle j'aimerais éviter de recourir à la physique quantique pour expliquer la boîte noire de l'intuition, c'est qu'il est dangereux, dans le cadre d'un travail de philosophie spéculative, d'appuyer des conclusions sur le fonctionnement empirique du cerveau. La force de la philosophie consiste en effet à discuter les relations nécessaires existant entre les concepts ; et non des mécanismes physiologiques contingents. Au vu de la complexité de notre organe cérébral, auquel même les experts en neurosciences reconnaissent encore bien des mystères, il est peu probable que la philosophie, en partant des idées universelles, puissent arriver à des conclusions précises.

D'autre part, et il s'agit là de la deuxième raison, on aurait beau jeu de substituer un mystère par un autre. Dire de cet oracle, de cette boîte noire énigmatique, qu'elle est le produit de phénomènes quantiques, sans donner à ces dernier toute la validité d'une théorie scientifique, c'est quitter le terrain de la philosophie spéculative pour s'aventurer, très maladroitement, sur celui de la science. Loin de participer à crédibiliser la philosophie, cela nous pousse, bien au contraire, à reconnaître qu'on a pas particulièrement avancé dans l'élucidation du mystère qu'on s'est chargé de révéler sous la pleine lumière du jour. Comme disait Maurice Maeterlinck, prix Nobel de littérature : « Il est rare qu'un mystère disparaisse ; il ne fait d'ordinaire que changer de place. »

Pour finir sur une troisième raison, autrement plus simple, pour laquelle je ne souhaite pas recourir à la physique quantique, notons simplement ceci : la rhétorique argumentative gagne à envisager toujours le cas le plus pessimiste. Plutôt donc que de s'aventurer sur des explications bancales à propos du fonctionnement explicite de notre boîte noire, acceptons avec mes détracteurs — ces lecteurs que l'on s'imagine toujours férus d'objections perverses — que cette boîte noire de l'intuition ne puisse pas recevoir d'élucidation ultime quant à son fonctionnement interne. Demandons-nous, sous cette hypothèse, si cela invalide les conceptions que j'ai défendue jusqu'à présent — et dans quelle mesure.

12.1 L'Insoutenable unité du corps et de l'esprit

J'aimerais prendre au sérieux l'idée que le modèle de l'esprit soit donné par une machine de Turing transfinie, tandis que le modèle du cerveau, lui, soit donnée par un mécanisme fini. À priori, ces deux affirmations sont contradictoires. Elles ne le sont cependant que si on admette que le modèle de l'esprit soit le modèle de cerveau, et réciproquement. Pourquoi faudrait-il que la façon dont fonctionne notre esprit puisse se réduire, dans ses termes les plus fondamentaux, à notre compréhension du cerveau ? Quelle autre alternative existe-t-il à cette posture philosophique ? Comment penser le rapport existant entre l'esprit (un modèle infini) et le cerveau (un modèle fini) ?

Dans ce chapitre, j'aimerais défendre un *mystérianisme de perspective*. Pour rappel, le mystérianisme consiste à affirmer qu'il existe des mystères, dans la nature, que l'esprit humain n'est pas en mesure d'élucider. On a vu que le philosophe McGinn, par exemple, soutient que la conscience fait partie de ce mystère (voir section [9.4.2](#)). J'aimerais marcher sur ses brisées, mais à quelques distinctions près. Je ne dis pas que l'esprit humain, en soi, est inexplicable. Ce qui est inexplicable, d'après moi, c'est le lien existant entre le modèle de l'esprit et le modèle du cerveau humain.

Précisons cette idée. Tout d'abord, j'affirme qu'il est possible (1) de

modéliser l'esprit au moyen de théories physiologiques décrivant le fonctionnement des cellules du cerveau ; c'est ce que font les neurosciences. Mais j'affirme également qu'il est possible (2) de modéliser l'esprit au moyen d'une théorie psychologique introspective. Cette théorie psychologique ne consiste pas à expliquer le fonctionnement du cerveau à l'aide de boîtes noires ou de lois quantiques, mais à rendre compte de la dynamique interne du psychisme depuis un point de vue subjectif — et pour ainsi dire phénoménologique. On trouve des exemples de cette approche dans la philosophie de Husserl, qui étudie les principes d'après lesquels des objets se constituent dans la conscience. On peut également penser à la philosophie kantienne, qui n'est pas fondamentalement éloignée d'une théorie structurelle de l'esprit humain. On peut aussi songer à la topique freudienne, qui tente de donner sens à des concepts tels que le Moi, la pulsion, l'angoisse ou le sentiment de culpabilité. Ces concepts n'ont pas de substrat physiologiques clairement identifiables. Les variables de base, ce ne sont pas des réactions biochimiques, mais des vécus de conscience : désires, représentations, idées, pensées, émotions, etc.^[1]

Ce qui demeure mystérieux, même en présence de ces modèles (1) physiologiques et (2) psychologiques, ce n'est pas l'esprit humain en tant que tel, mais la relation que notre esprit entretient avec le corps — et plus généralement le cerveau. On comprend bien que le psychisme possède des lois, des pulsions, des dynamiques ; l'expérience de vie et l'introspection nous permet de les entrevoir. Ce qu'on ne comprend guère, en revanche, c'est la relation existant entre notre cerveau et les vécus de conscience (sensation de froid, désir, la pulsion de mort...). Existe-t-il un lieu de causalité entre nos pensées et le fonctionnement de nos cellules nerveuses ? Ce lien est-il unidirectionnel ou bidirectionnel ? Comment expliquer qu'une chose aussi immatérielle que la pensée soit causée par des changements au sein de la matière ?

La perception vécue à la première personne ne semble pas du même ordre que l'observation de notre cerveau à la troisième personne. C'est pourquoi il existe un sentiment incurable d'étrangeté lorsque nos yeux, d'aventure, se perdent dans la contemplation de nos bras, de notre tête, de nos lèvres contrariées par les mystères obtus de la nature. Il ne semblerait pas que le sentiment d'être vivant coïncide avec la représentation, tout à fait extérieure et objective, de ce corps que nous sommes.

Le mystérianisme que je vais défendre dans ce chapitre consistera à penser que si la relation entre le corps et l'esprit est énigmatique, c'est parce qu'il s'agit d'une seule et même réalité — dont nous sommes incapables de comprendre l'unité fondamentale. Nous en sommes incapables en raison de la nature même de notre intelligence, et non pas par manque de savoir. Cette

1. Il est à noter que Freud, neurologue de formation, voulait donner à sa théorie un substrat physiologique précis. Dans son essai de jeunesse, intitulé *Esquisse d'une psychologie scientifique*, il modélise le cerveau par le biais de principes inspirés de la thermodynamique. Force est de reconnaître que sa théorie s'est peu à peu détachée d'une description purement naturaliste.

intelligence, en effet, est située dans le temps et l'espace. Elle est un point de vue subjectif sur le monde. Tout ce qu'elle voit, ce sont des phénomènes au sein de cette structure spatio-temporel. Or elle cherche à voir, non pas les phénomènes apparaissant au sein de cette structure, mais les contours de la structure elle-même. Le problème de la conscience, ainsi compris, n'est-il pas aussi difficile que d'observer le temps sans aucun événement ? Ou d'observer l'Univers en l'absence de matière ? Le temps ne se manifeste qu'au travers d'événements. L'Espace ne se donne à voir qu'au travers de corps étendus. Sans corps, sans événements, il n'y a rien de concevable.

Pour vous donner une métaphore, on pourrait dire que l'unité du corps (une sorte d'objet spatial) et de l'esprit (une sorte d'événement temporel) se présente à nous de la même façon qu'un objet cylindrique se présenterait à un être à deux dimensions. Pour cet être, le cylindre apparaîtrait tantôt sous la forme d'un cercle, tantôt sous la forme d'un rectangle, selon qu'il soit projeté d'après une coupe longitudinale ou transversale. Autrement dit, notre être bidimensionnel verrait deux objets différents, un cercle et un rectangle, selon le point de vue adopté. Il n'aurait aucune raison de croire, *a priori*, qu'il s'agit de deux perspectives portant sur un même objet de la troisième dimension.

Tout comme cet être bidimensionnel, nous voyons la réalité tantôt sous le prisme de l'esprit, tantôt sous le prisme de la matière, et nous soupçonnons qu'il existe un lien entre ces deux perspectives. Nous faisons donc l'hypothèse (1) que l'esprit et la matière sont deux réalités différentes et (2) que ces réalités présentent un point de rencontre dans le cerveau. Nous pensons en outre que (3) la réalité de l'esprit doit pouvoir s'expliquer uniquement par celle de la matière, si bien que les termes se rapportant à esprit (pensées, idées, sensations, douleur...) ne soient finalement que des phénomènes physiologiques.

Cette vue que je viens de décrire est celle du matérialisme réductionnisme. Dans ce chapitre, j'aimerais suggérer tout le contraire ; à savoir (1) que l'esprit et la matière sont une seule réalité ni purement mentale, ni purement spatiale, mais (2) qu'il n'existe aucun point de rencontre entre l'esprit et la matière et (3) que la réalité ne peut être entièrement expliquée par l'esprit, ni entièrement expliqué par la matière. Je défendrai que notre compréhension du monde exige de passer par l'une et par l'autre de ces perspectives. Il existe des problèmes qui reçoivent une réponses adéquats sous l'angle de l'esprit (ai-je mal ? suis-je heureux ? suis-je conscient en cet instant t ?) et d'autres qui exigent de prendre l'angle de vue de la matière (l'Univers est-il fini ? comment produire de l'énergie ?)

Pour rendre ma position philosophique crédible, je vais d'abord donner des arguments montrant que le mystère de la relation entre l'esprit et le corps, et particulièrement le cerveau et la conscience, présente un paradoxe logique insoluble. Chaque fois que nous mettons en parallèle une expérience de conscience subjective (perception, sensation, idée) avec un phénomène

physiologique (excitation neuronale), en supposant qu'il existe une causation entre l'apparition d'un signal électrique et la naissance d'une pensée, nous commettons une erreur de jugement. Cette erreur est ce que j'appelle un chiasme métaphysique ; elle consiste à mélanger des termes qui n'ont, entre eux, aucun rapport. Plus précisément, je montrerai que le paradoxe auquel on se heurte est un paradoxe d'autoréférence. Je distinguerai aussi deux types de chiasmes métaphysiques : les chiasmes dits *internalistes* et les chiasmes *externalistes*.

Je ne me contenterai toutefois pas de séparer, d'un coup de bâton au milieu de la mer rouge, ce qui a trait à l'esprit et ce qui a trait à la matière. Je montrerai que, bien qu'on ne puisse pas penser le rapport entre la conscience et la matière, il est impossible de concevoir la conscience sans faire implicitement référence à la matière ; et il est tout aussi bien impossible de concevoir la matière sans être renvoyé implicitement du côté de la conscience. Autrement dit, je tenterai de suggérer que le corps et l'esprit sont comme le jour et la nuit. Sans lumière, il est impossible de penser ce qu'est une ombre ; mais sans ombre, point de lumière. On ne peut penser l'objectivité de la matière qu'en se référant à la subjectivité de la conscience ; et réciproquement.

Cette collusion entre conscience et matière, cette intrication logique fondamentale des concepts, me conduira à poser comme hypothèse que le corps et l'esprit ne sont que les deux facettes d'une même réalité substantielle — les deux visages d'une déesse intemporelle. On ne peut pas les séparer, puisqu'ils ne forment qu'un seul tout ; mais on ne peut pas non plus expliquer leur liaison. La malédiction de la raison humaine consiste à abstraire si bien les termes les uns des autres, pour leur donner des définitions indépendantes, qu'elle se trouve dans l'incapacité de penser l'unité dont ces termes sont issus préalablement.

Ce mystérianisme dont j'ai esquissé le portrait est ce que j'appellerai un *mystérianisme perspectiviste*. La stratégie de ce chapitre est plurielle. Dans un premier temps, je poserai quelques hypothèses à propos de la conscience et je montrerai en quoi ces hypothèses conduisent au fameux paradoxe de la conscience. Ensuite, je montrerai que ce problème peut être contourné si l'on admet que le corps et l'esprit forment une unité cohérente, mais pas totalement explicable. Dans les chapitres suivants, je tenterai alors de montrer que ma vision philosophique est cohérente avec les théories contemporaines du cerveau et de la conscience.

12.2 Hypothèse 1 : Conscience vive et conscience morte

Du point de vue purement intime et introspectif, la seule chose indubitable pour un philosophe rationaliste est d'être, à un instant donné,

conscience de quelque chose plutôt que rien. Le monde tangible qui se déploie sous le tâtonnement de ses doigts, de ses yeux, de sa langue, n'existe que dans la conscience qu'il en a. L'idée même d'être un sujet percevant le monde, un corps au sein de la matière, une personnalité pensante dotée de son histoire, de sa mémoire, de son caractère immuable, semble pouvoir être remise en cause. Seul existe l'instant dans sa plénitude.

Du point de vue de l'expérience scientifique en revanche, il semblerait que la certitude première soit l'existence d'un monde matériel, fait de hadrons, d'ondes et d'énergie, au sein desquels il est difficile d'expliquer comment émerge le phénomène de la conscience proprement dit. Lorsque je regarde le ciel sempiternel, que je me représente, un à un, l'infinité des mondes piqués sur la grande toile de la Voie lactée, comme tout autant d'étoiles en perdition dans le vide glacé de l'abîme, je ne puis m'empêcher de ressentir un curieux vague à l'âme. L'Univers me semble effroyablement plus réel que la conscience que j'en ai. Il est peuplé de sons, de couleurs, de fragrances enivrantes, qui ne sont rien d'autre que des particules quantiques en déplacement stochastique. Ce que je prends pour un authentique phénomène de conscience (la contemplation du ciel) n'est qu'un peu de bruit et de fureur (des réactions neurologiques dans le cerveau).

L'on pourrait donc penser, en guise de première hypothèse, que la conscience peut être définie de deux façons distinctes :

- En un sens *naturaliste*, qui consiste à la regarder comme une boîte noire dont il convient de simuler, de reproduire ou de prédire les comportements. Les données récoltées par suite d'expériences donnent de nombreuses informations sur les spécificités de la conscience en tant que fonction de l'organisme, mais non sur sa nature en tant que vécu. À ce titre, la démarche pourrait être qualifiée d'externe et d'objective. Externe, car elle porte sur les manifestations représentables et quantifiables de la conscience. Objective, car le point de vue adopté est un point de vue à la troisième personne. Sur le plan épistémique, il est présupposé que le résultat de l'expérience n'est pas teinté de la subjectivité de l'observateur : l'objet est prétendument donné à voir tel qu'il est, et non tel qu'il est vu. Sous cette acception tombent certaines définitions, ou caractéristiques de la conscience, que l'on trouve dans la littérature, notamment sous des termes comme « *conscious creature*^[2] » « *access consciousness* », « *monitoring consciousness*^[3] », mais aussi dans les modèles computationnels de la conscience, qu'ils

2. VAN GULICK R., *What would count as explaining consciousness ?* In T. Metzinger, ed. *Conscious Experience*. Paderborn : Ferdinand Schöningh, 1995.

3. Block Ned, *A Theory of Consciousness in The Nature of Consciousness : Philosophical Debates*, Cambridge : MIT Press 1997

soient basés sur l'attention^[4] ou sur l'intégration de l'information^[5]. Ces modèles se basent souvent sur une impasse de raisonnement comparable au problème de l'induction relevé par Hume dans son *Traité sur l'Entendement Humain*. Avec cette acception de la conscience, purement externe et descriptive, la science a parfois tendance à réduire la vie mentale à un ensemble de processus biologiques à propos de la matière morte. Pour cette raison, la conscience naturaliste sera parfois désignée, de façon plus caustique, sous le nom de *conscience morte*. Cette conscience, du reste, n'a d'autre viser que d'offrir un modèle du cerveau.

- En un sens *phénoménologique*, qui consiste à la regarder comme un principe de connaissance. C'est l'approche qui est adoptée dans le *cogito* cartésien, mais aussi dans les expériences spirituelles. Par expériences spirituelles, nous désignons toutes les pratiques visant à éveiller la conscience à son monde intérieur, ou plus précisément, à l'idée que la conscience forme une unité totalisante en dehors de laquelle rien ne peut être conçu. Cette définition de la spiritualité n'est pas très éloignée de celle donnée par Foucault dans *L'Herméneutique du Sujet*. Dans la mesure en effet où le sujet connaissant est aussi son propre objet de connaissance, la recherche de vérité s'inscrit dans une démarche de transformation de soi. Toute connaissance acquise sur la conscience est une connaissance acquise sur ce que l'on est (ou ce que l'on est pas). Il s'agit donc d'une démarche interne et subjective. Interne, car elle porte sur la conscience telle qu'elle est en son intimité. Subjective, car le point de vue adopté ici n'est pas celui de Dieu, mais est un point de vue à la première personne. Sur le plan épistémique, il est présupposé que le sujet n'a pas besoin de faire référence à autre chose qu'à sa propre expérience pour parler de son vécu : l'on a tendance à oublier que le langage, la culture et les représentations objectives puissent par exemple jouer le rôle de filtres ou de biais dans la compréhension de lui-même. Il s'ensuit que la conscience est trop souvent vue comme un état de fait, où les phénomènes se donnent de façon immédiate et représentationnelle. Sous cette acception tombent certaines définitions, ou caractéristiques de la conscience, que l'on trouve dans la littérature, notamment sous des termes comme « qualia », « phenomenal structure » et « subjectivity^[6] », "self-consciousness" et "phenomenal consciousness" (BLOCK NED, *op. cit.*), « introspective knowledge » et « homogeneity^[7] ou en-

4. TAYLOR, *Paying attention to consciousness*, Progress in Neurobiology, 2008.

5. La Théorie de l'Information Intégrée, fondée par Giulio Tononi et Gerald Edelman, fera l'objet du chapitre ^[15].

6. VAN GULICK R., *op. cit.*

7. LYCAN WILLIAM G., *Consciousness*, First published October 1, 1987.

core « first person perspective⁸ ». Dans la mesure où cette définition de la conscience tient compte du vécu dans son apparition immédiate, mais non des structures sous-jacentes à ce vécu, j'appellerai cette conscience la *conscience vive*. Cette conscience vise à donner un modèle de l'esprit humain.

Au vu de ce qui vient d'être dit, tout contenu mental présente une part *d'observation* et une part *d'introspection*. L'introspection envisage ce contenu mental en tant que phénomène de conscience, en faisant abstraction de l'existence d'une réalité au-delà des sens ; l'observation au contraire retranche tout ce qu'il y a de subjectif dans un contenu mental pour ne conserver que les objets qui sont visés. Si l'introspection retient de la pensée d'une pomme que « Je vois une pomme », l'observation pour sa part ne retient que ceci : « Il y a une pomme ». Dans le premier cas, la connaissance évolue au sein de la conscience vive ; dans le second cas, elle vise une réalité au-delà de l'expérience elle-même. Pour l'exprimer en termes husserliens, l'introspection représente la dimension phénoménale de la conscience (c'est-à-dire le fait même de sentir, de penser, se vivre,) tandis que l'observation représente la dimension naturaliste (l'objet intentionnel sur lequel se porte la pensée).

L'existence d'une vie introspective, d'un vécu subjectif à la première personne, a souvent été pris comme pierre de touche de la philosophie. En témoigne notamment le cogito cartésien, la certitude sensible de Hegel, le Je transcendantal de Kant ou les monades leibniziennes. Elle présente néanmoins une difficulté dont il est important de comprendre les tenants et les aboutissants. Lorsque l'on part de l'introspection, c'est-à-dire de la perspective en Je, on a sous les yeux le phénomène vif de la conscience, mais il demeure le problème de déterminer les limites de cette conscience. Au même titre que le poisson, plongé dans l'Océan, se demande ce que c'est que l'eau dans laquelle il baigne et quels en sont les contours, la philosophie est mise en demeure de déterminer le bien fondé du concept de « Je ». Au cours de cette thèse, je ferai l'hypothèse que le problème du « Je » est un problème de représentation, qui fait de la conscience un espace ouvert, sans clôtures nettement définies.

12.3 Hypothèse 2 : Le Paradoxe de la conscience

La deuxième hypothèse que je souhaiterais introduire à propos de la conscience concerne l'articulation entre l'approche naturaliste (la conscience morte) et phénoménologique (la conscience vive). Avant d'introduire cette hypothèse, j'aimerais rappeler à l'esprit du lecteur des concepts de topologie

8. NAGEL THOMAS, *What Is It Like to Be a Bat ?*, The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.

qui devraient, je crois, l'aider à mieux cerner ce qui va suivre.

En topologie, on distingue les ensembles *ouverts* des ensembles *fermés*. Un « ensemble ouvert », ou plus simplement un « ouvert », est une zone de l'espace délimitée par un certain contour, mais qui exclut les points constituant ce contour. Un exemple concret d'un tel espace peut être obtenu en considérant un disque dont on enlève le pourtour. Un ensemble fermé en revanche est une zone de l'espace incluant le contour.

En gardant ces distinctions à l'esprit, ma deuxième hypothèse consistera à défendre que la conscience est un espace *ouvert* : bien que toute la réalité soit dans la conscience, les contours de la conscience elle-même ne sont pas représentables dans la conscience. Comment en effet pourrait-on avoir une représentation finie et délimitée de la capacité, infinie et sans cesse renouvelée, de se faire des représentations ?

Cette difficulté de conscientiser la conscience peut recevoir une analogie. Supposons qu'un individu ait grandi toute sa vie dans une pièce dépourvue de toute surface réfléchissante. Il est capable de voir le monde autour de lui : les murs, les tables, les chaises, etc. Il peut également contempler ses mains, son corps et — avec un peu de chance — le bout de son nez rougi par une trop grande quantité d'alcool, mais il n'a jamais vu son propre visage, ni la couleur de ses yeux. Comment un tel homme peut-il expliquer le phénomène de la vision ?

Notre individu hypothétique comprendrait certes que la vision, en tant qu'expérience quotidienne, doit avoir une origine en lui. Il pourrait même affirmer que cette origine se trouve quelque part sur son visage, mais, bien qu'il puisse fermer ses paupières, palper ses yeux du bout des doigts, il ne serait pas capable de voir ses yeux de ses propres yeux. Autrement dit, la vision serait pour lui une évidence sensitive, mais pas une évidence d'ordre visuel. Il aurait conscience de son propre œil comme de quelque chose d'abstrait, qu'il peut toucher avec ses mains, qu'il peut sentir en lui, mais qu'il ne peut pas voir.

Que devient cette analogie lorsqu'on la rapporte, non pas à l'œil de la vision, mais à la conscience ? Au même titre que l'œil ne tombe pas lui-même sous le sens de la vue, il se pourrait que la conscience ne soit pas, en tant qu'expérience, appréhendable par les ressources de notre conscience. Chaque fois que je me représente moi-même en tant que conscience, que je me dis « Je suis un tel », « Je suis celui qui pense », « Je suis cogito », « Je suis le sujet de toute expérience » ou « je suis un cerveau dans le monde » ; chaque fois que je me livre à mon auto-contemplation, j'ai sous les yeux une image, une émotion, une sensation de moi-même. Cette dernière est un objet de ma pensée comme tant d'autres et ne se distingue pas en cela de la représentation d'une table, d'un livre ou d'un singe. Cette image de moi n'est pas ma conscience objective en tant que telle, mais une simple représentation subjective contenue dans mon expérience de conscience.

Cet apparent paradoxe est peut-être la raison fondamentale pour laquelle

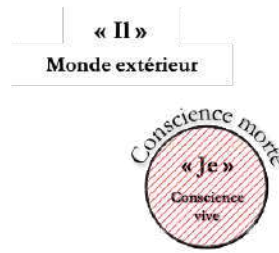


FIGURE 12.1 – Le cercle (aire hachurée et contour) symbolise la conscience au sens large. Cette dernière se décompose en deux parties : la conscience vive (aire hachurée) et la conscience morte (le contour du cercle). La première est composée de l'ensemble des contenus conscients possibles et imaginables ; la seconde est le fait même d'être conscient, c'est-à-dire de contenir des contenus conscients. On a alors deux approches : (a) L'approche introspective part du point de vue en « Je » et tente de décrire la conscience morte en des termes phénoménologiques ; (b) L'approche scientifique part du point de vue en « Il » et tente de décrire la conscience vive en des termes naturalistes. Le monde extérieur du « Il » tout comme le monde intérieur du « Je » sont des espaces ouverts.

le monde conscient semble peuplé d'objets qui sont en moi mais qui ne sont pas *moi* ? La grande idée de cette deuxième hypothèse consiste à penser que la conscience, contrairement aux objets usuels, ne peut pas être représentée. Elle peut soit être *vécue* subjectivement comme un « Je » (je suis conscience), soit *conçue* abstraitement et objectivement comme un « Il » (il y a de la conscience).

Dans le premier cas, la conscience est alors un je-ne-sais-quoi dont on ne peut rien dire. Elle se manifeste, mais dès l'instant où elle apparaît, elle n'est déjà plus. Dans le second cas, elle n'est rien qu'une idée abstraite, qu'on ne peut pas se représenter concrètement. Il en va un peu de cette conscience abstraite comme de la proposition « L'Homo sapiens est l'être qui pense qu'il pense ». On comprend bien que l'Homo sapiens pense au fait qu'il pense, mais on ne sait pas à quoi il pense concrètement. Penser au fait de penser ne semble pas aussi visuel que penser à la pomme appétissante qui se trouve au milieu de mon jardin.

12.3.1 La Conscience est irréprésentable

Conformément à notre première hypothèse, le mot « conscience » désignerait donc deux concepts distincts. D'une part, il s'agirait de la succession de toutes les expériences possibles et, d'autre part, de l'idée non représentable du « Je » qui renferme ces expériences.

La première dimension est ce que j'ai appelé la *conscience phénoménologique*. Elle porte toujours sur un contenu précis : une perception, une idée,

une émotion.... À ce titre, on peut dire que la conscience phénoménologique qu'elle est une conscience représentationnelle ou, pour employer les termes de Husserl, qu'il s'agit d'une conscience intentionnelle. Cette conscience est toujours conscience de quelque chose. Il n'est pas possible d'être conscient de rien, pas plus qu'on ne peut se représenter le rien.

La seconde conscience, la *conscience naturaliste*, existe quant à elle au-delà de tout contenu mental déterminé : elle est le contenant au sein duquel se déploient les phénomènes conscients. C'est donc une conscience non représentationnelle. Cette irréprésentationnalité est d'ailleurs parfaitement résumée par l'aphorisme sartrien : « on *vit* l'intériorité de la conscience, mais on ne la contemple pas⁹. » On pourrait dire également de la conscience morte qu'elle est une conscience hyperintentionnelle car elle se tient au-delà de toutes les visées intentionnelles possibles et imaginables.

Pour récapituler, la deuxième hypothèse consiste à penser que le difficile problème de la conscience, celui qui vise à saisir le lien entre le corps et l'esprit, est un problème de représentation. Plus précisément, je voudrais défendre que, à l'origine de la visible impossibilité que nous éprouvons à localiser la conscience dans le cerveau — et plus généralement, dans l'espace matériel — se trouve l'impossibilité, pour un esprit fini tel que le nôtre, de se faire une représentation de l'infini en acte. C'est-à-dire que nous ne pouvons pas circonscrire dans une pensée délimitée la fécondité illimitée de la conscience, entendue comme le pouvoir de créer sans cesse de nouveaux contenus mentaux.

Remarquons que lorsque je dis que la conscience est « irréprésentationnelle », je ne veux pas dire qu'elle soit « irréprésentable » *stricto sensu*. Il va de soi que la conscience morte, celle dont parle la science, peut être représentée dans le langage mathématique et neuroscientifique. C'est un objet formellement défini, doté de processus neuronaux, de fonctions mathématiques, régi par des hypothèses vérifiables et conduisant, dans le meilleurs des cas, à des prédictions. Ces propriétés, représentables dans l'esprit du scientifique, ne sont cependant pas *l'expérience de conscience* visée par le scientifique. Il y a donc représentation symbolique d'une conscience qui, par nature, se tient au-delà des représentations conscientes. D'une certaine manière, on pourrait dire que ce qu'exprime la science à propos de la conscience, elle l'exprime en un sens analogique.

Quand Antonio Damasio ou Gerald Edelman parlent de l'excitation d'un groupe de neurones dans la région du cortex somesthésique, ils veulent désigner par là l'existence d'une certaine proprioception. Ils savent bien entendu que le langage utilisé pour décrire cette sensation n'est pas la sensation elle-même. Leurs théories scientifiques convoquent des représentations auxquelles

9. La citation exacte est un peu moins élégante, et je me suis permis de la simplifier pour les besoins de l'argumentation : , dans SARTRE, J.-P., *La Transcendance de l'Ego : Esquisse d'une description phénoménologique*.

tout le monde à accès dans sa conscience : la représentation de neurones, d'activités synaptiques, de figures mathématiques, etc. Mais aucune de ces images ne sont, en tant qu'images, la conscience morte proprement dite : ce sont tout au plus des contenus de la conscience vive parmi d'autres. Il ne viendrait à l'idée de personne de dire d'un objet jaune qu'il soit le jaune à proprement parler ? Ou du dessin d'un arbre qu'il soit le concept de l'arbre ? La science utilise des représentations conscientes qui désignent, en réalité, un phénomène de conscience situé au-delà des représentations conscientes.

Bien sûr, la conscience n'est pas le seul objet scientifique à se tenir dans un rapport de représentation. Lorsque la physique théorique parle de particules fondamentales, les symboles e ou p *représentent* ces particules — mais tout le monde sait que les symboles ne sont pas eux-mêmes des électrons ou des protons. Il y a toutefois une différence méthodologique fondamentale entre les objets de la physique et cet objet singulier qu'est la conscience.

Lorsque je dis qu'un électron a telle charge électrique, telle masse, tel spin, je ne fais jamais que parler d'une réalité existant « en dehors » de la conscience. Il est donc tenu pour acquis que les images mentales subjectives que je me fait de l'électron ne sont pas des électrons proprement dit. D'après l'idéal scientifique, je devrais même mettre toutes mes conceptions personnelles entre parenthèses pour comprendre le plus adéquatement possible mon objet d'études.

Or cette neutralité ne tient pas dans le cas de la conscience. Si je dis de la conscience qu'elle est « un phénomène de réentrance des signaux cérébraux envoyés par le système thalamo-cortical¹⁰ », je ne peux pas mettre m'empêcher de les comparer aux contenus subjectifs de la conscience, puisque, par définition, la conscience est l'ensemble des contenus subjectifs.

Ce paradoxe de la représentation est peut-être ce qui explique que les individus, lorsqu'ils lisent une nouvelle définition scientifique de la conscience, éprouvent parfois la sensation que cette définition ne correspond pas à leur vécu ? Ils objecteront que la conscience n'est pas un mécanisme, mais qu'il y a bien un fantôme dans la machine : une « âme », un « esprit » ou un « principe vital ». Ces résistance du sens commun à l'endroit de la démarque scientifique doivent être prises au sérieux. Elles suggèrent que réduire les représentations de la conscience morte à la conscience proprement vivante est une opération douteuse, que je désignerai sous le nom de *chiasme métaphysique*. Je vais montrer qu'il en existe de deux sortes.

12.3.2 Chiasmes externalistes et internalistes

L'étude de la conscience morte est pertinente lorsqu'il s'agit de décrire ou de prédire le comportement d'un sujet conscient dans un espace d'objectivité. L'étude de la conscience vive, quant à elle, est pertinente lorsqu'il

10. Cette définition, comme on le verra plus loin, est celle de Gerald Edelman.

s'agit de comprendre ou de faire sens de ce que nous sommes en tant que totalité perceptive. Mais dès lors que l'on tente d'adresser des questions relevant de la conscience morte à partir du point de vue à la première personne, ou inversement lorsque l'on tente de pénétrer la conscience vive à partir du point de vue à la troisième personne, l'on se heurte à des chiasmes méthodologiques :

- **Chiasme externaliste** : Les écueils d'une démarche en « Il » se font apercevoir lorsque la science, cessant de prendre pour objet d'étude les manifestations externes de la conscience, tente d'en appréhender la vie intime. Cette approche est problématique, car il y a comme une pétition de principe. Il est en effet postulé que les mesures obtenues au sein d'un protocole expérimental se réfèrent à l'intériorité du sujet, alors qu'elles ne sont en fait que le fruit d'une mise en correspondance entre ce qui est observé sur le plan strictement factuel à l'aide, par exemple, d'un encéphalogramme, et ce que le sujet prétend ressentir au moment où ces mesures sont prises. Plutôt que de lier par une relation de cause à effet un état interne de la conscience avec une manifestation externe, on ne fait, en réalité, que mettre en correspondance deux espèces différentes de manifestations externes : les mesures physiologiques et les déclarations orales. Cette difficulté ne relève pas de l'imperfection des outils de mesure, mais de ceci qu'il est impossible de "pénétrer" dans la tête de quelqu'un d'autre et de ressentir ce qu'il sent. C'est une impasse de raisonnement comparable au problème de l'induction relevé par Hume dans son *Traité sur l'Entendement Humain*, puisque la conclusion qui est visée ne peut pas être prouvée de façon strictement rationnelle.
- **Chiasme internaliste** : Les écueils d'une démarche en « Je » se font apercevoir au contraire lorsque, partant d'un point de vue à la première personne, l'on tente de s'exprimer sur une réalité extérieure à la conscience vive. Ce geste est impossible puisque, au sein de cette attitude, tout ce qui est intelligible fait partie de la conscience à part entière. Le problème de la conscience vive, pour le dire autrement, c'est qu'elle se présente comme un contenant qui englobe toute la phénoménalité, mais dont il est impossible d'apercevoir les contours. Nous sommes placés, dans notre tentative de décrire cette conscience, dans une position comparable à celle d'un poisson qui entreprendrait de définir ce que c'est que de l'eau. Dans la mesure où l'eau est son milieu ambiant, il lui est impossible de sortir de ce milieu pour le décrire depuis l'extérieur. De la même manière, il est impossible de sortir de la conscience vive et de l'isoler pour la faire entrer dans une catégorie intelligible, sauf à abandonner le point de vue à la première personne et à se priver, dans le même temps, de l'expérience

de conscience elle-même. Il est donc aporétique, au sein d'une démarche en « Je », de poser la question de savoir ce que c'est que ce « Je » qui pense : une substance ? un objet du monde ? Le chiasme métaphysique internaliste est comparable, non pas au problème de l'induction, mais aux paradoxes de l'autoréférence. Elle suppose naïvement que le sujet, par introspection, a un accès direct à ce qu'il est lui-même.

Ces définitions ayant été posées, nous allons défendre, tout au long de ces pages, l'idée qu'il n'existe pas de conscience subjective pure ni de conscience objective pure, mais que les recherches en sciences cognitives se trouvent toujours dans une démarche croisée, et donc méthodologiquement problématique. Pour ce faire, il s'agira dans un premier temps de montrer, en étudiant les structures épistémiques des expériences scientifiques, que chaque fois qu'une théorie prétend dire quelque chose de la conscience morte pure, elle n'est pas tout à fait dans une démarche d'extériorité, mais dissimule en fait un chiasme externaliste, si bien que la seule conclusion possible, si elle voulait être rigoureuse, consisterait à dire que la conscience vive n'existe pas¹¹.

Ensuite, nous allons suggérer, en partant de l'étude du cogito cartésien, que chaque fois que la conscience vive prétend parler d'elle-même comme d'un objet dans le monde, elle produit un chiasme internaliste, si bien que la seule conclusion possible, à ses yeux, consisterait à dire que tout n'est que de la conscience ; ou, ce qui revient au même, que la conscience n'est qu'un tout indifférencié. Dans le premier cas, la conscience objective pure, forcée de se réfugier dans ses derniers retranchements, s'efface devant la loupe de l'expérimentateur comme par implosion. Dans le deuxième cas, la conscience subjective pure, incapable de se voir elle-même, s'évanouit comme par dilatation.

Dans la mesure où il n'existe pas de conscience subjective pure et de conscience objective pure, il va être nécessaire, dans un second temps, de poser la question de savoir d'où peut bien venir l'idée d'une extériorité de la conscience vive ou d'une intériorité de la conscience morte ? C'est une chose de dire que ces questions sont dépourvues de sens, c'en est une autre de dire d'où elles proviennent. Si en effet ces interrogations étaient réellement contraires à la pensée rationnelle, l'on serait en droit de croire qu'elles n'auraient pas même pu être conçues. Or il se trouve qu'elles sont bien présentes en nous, et nourrissent si bien les échafaudages du langage que celui-ci semble tout entier appuyé sur des mots s'y référant : objectif, subjectif, intérieur,

11. L'éliminativisme est défendu par Patricia Churchland (1986) et Paul Churchland (1981) et par les tenants de l'illusionnisme. Cette idée encore plus radicale consiste à dire qu'un matérialisme épuré de ses contradictions nécessite de voir ce que j'ai appelé la conscience vive comme une illusion. Se référer notamment à l'ouvrage *Conscience et matière* de mon ex-collègue FRANÇOIS KAMMERER.

extérieur, corps, matière, esprit, âme, etc.

Ceci met la philosophie en position d'entreprendre une recherche sur la genèse de ces concepts. Elle le fait en vue d'expliquer, non pas ce que c'est que l'extériorité ou l'intériorité, mais ce que c'est que l'idée même de l'extériorité et de l'intériorité. À ce titre, la démarche peut être qualifiée de phénoménologique et de transcendantale. Phénoménologique, car elle évolue au sein même du phénomène, mais qu'elle tente de dépasser la simple immédiateté perceptive de l'expérience vécue au moyen d'un langage dont la signification renvoie à des objectivités, ou plus exactement des visées intentionnelles de la pensée. Transcendantale, car la question qui est adressée, c'est quelles sont les conditions qui rendent possible l'idée d'une intériorité de l'objectivité et d'une extériorité de la subjectivité.

12.3.3 Liens avec les paradoxes logiques

J'ai décrit en quoi le rapport entre la conscience vive et la conscience morte était paradoxal, mais sans définir de façon formelle la nature de ce paradoxe. Je me suis contenté de parler par analogie, en prenant l'exemple d'un homme qui chercherait à voir son propre œil. Il est toutefois légitime de se demander si le problème de représentation de la conscience est de la même nature que les paradoxes logiques connus par ailleurs, tels que le paradoxe du Barbier (section 5.2.1) ou le paradoxe de Cantor (section 5.2.2).

Non content de clarifier mon propos, cette question que je soulève est également motivée par un problème plus fondamental : comment serait-il possible de penser les paradoxes logiques (en ce compris le paradoxe du Barbier) si la conscience n'était pas elle-même de nature paradoxale ? Je vais m'expliquer là-dessus.

Par définition, la conscience vive regroupe l'ensemble des phénomènes. Cela veut dire qu'aucune pensée, en dehors de la conscience, ne peut être conçue. Tous les savoirs objectifs que nous détenons à propos du monde, toute notre science, notre génie, notre art, les livres accumulés depuis des générations, existent au sein de la conscience humaine (quoiqu'il soit bon de préciser qu'ils n'existent pas forcément uniquement au sein de celle-ci. Mon point consiste à dire qu'ils existeront, en tout cas, dans la conscience.)

À titre d'exemple, notre connaissance mathématique de l'espace géométrique correspond à une expérience consciente, que nous pouvons reproduire en pensée, en nous imaginant des corps, et leur faisant subir des translations, des rotations. Il en va de même de la connaissance des nombres, des corps, des électrons.

Si notre conscience du monde n'avait pas la structure de l'espace euclidien, il nous serait impossible de nous représenter l'espace euclidien. C'est d'ailleurs parce que notre expérience consciente n'a pas la structure d'un espace à quatre dimensions que nous sommes incapables de nous représenter concrètement un hypercube ou une hypersphère à quatre dimensions. Ce

que je veux dire, en des termes certes kantien, c'est que les objets de notre savoir épousent les structure de notre esprit humain.

La question qui s'impose dès lors est la suivante : de quelle expérience de conscience viennent notre connaissance des paradoxes logiques ? Si la conscience n'avait rien de paradoxale et de contradictoire, comment expliquer que nous puissions penser des objets contradictoires ?

Cette ligne de pensée admet deux réponses possibles. Soit l'on dit, comme le pensait jadis le mathématicien Henri Poincaré, que les paradoxes logiques ne sont que des jeux de langage ; c'est-à-dire des choses abstraites et irréprésentables, qui n'ont pas d'existence au-delà de leur formulation linguistique. Cette opinion était majoritaire au début du vingtième siècle, lorsque la signification des paradoxes de Russell ou de Cantor n'était pas encore pleinement connue. Il est toutefois admis que ces paradoxes, loin d'être anecdotiques, ont des conséquences majeures sur notre conception des mathématiques (se référer à la section 5.2.5 pour de plus amples exemples).

Soit, et il s'agit de la seconde option, l'on admet au contraire que les paradoxes logiques correspondent à des réalités représentables de la conscience. Ce dernier cas de figure s'apparente au dialethéisme, qui admet l'existence d'objets paradoxaux, qui sont à la fois vrai et faux. Des défenseurs célèbres de cette position sont Hegel, Łukasiewicz ou Graham Priest.

Même en admettant le dialethéisme, il n'est toutefois pas tout à fait intuitif d'expliquer comment l'être humain pourrait se représenter un cercle carré. Dès que nous nous représentons un cercle-carré, une image nous vient à l'esprit ; mais une simple inspection suffit à saisir que cette image ne correspond ni à la définition d'un cercle, ni celle d'un carré. Il est, à vrai dire, impossible de se faire une représentation concrète, sensible, d'un objet contradictoire ; si bien qu'il est plus facile de dire que de tels objets n'existent pas. Notez que ce raisonnement est valable avec les autres types d'objets paradoxaux comme « l'ensemble de tous les ensembles » (paradoxe de Russell) ou « le cardinal de tous les cardinaux » (paradoxe de Cantor) ou « le plus petit entier non nommable en moins de dix-huit syllabes¹² » (paradoxe de Berry).

Il semblerait que les objets paradoxaux occupent une position intermédiaire entre des objets représentables et des objets qui n'existent pas : ce sont des réalités logiques sans représentation explicite. Nous pouvons les concevoir abstraitement, mais nous ne pouvons pas les sentir, les voir, les dessiner en imagination — ni même les penser de façon cohérente. Il faut donc admettre qu'il existe, dans la conscience, des objets abstraits non représentables. De tels objets occupent la frontière de la conscience. Ils sont dans la conscience, puisque nous pouvons les concevoir, mais en même temps ils n'y sont pas ; puisque nous ne pouvons pas nous les représenter.

Or quel est le seul objet de la conscience qui est concevable sans être

12. Ce paradoxe vient de ce que la phrase « le plus petit entier non nommable en moins de dix-huit syllabes » est elle-même une phrase de dix-huit syllabes.

représentable ? D'après notre seconde hypothèse, il ne peut s'agir que de la conscience elle-même. La frontière de la conscience vive, comme on l'a dit, c'est l'idée objective de la conscience. Il s'ensuit que l'expérience qui est à l'origine des objets contradictoires — ce qui rend possible notre théorie des paradoxes — c'est l'expérience de conscience.

J'entends déjà le lecteur m'objecter que la conscience n'est pas le seul objet non représentable. Les figures de l'espace en quatre dimensions, dirait-il, ont également cette propriété d'être des objets concevables, mais non représentables. Il en va de même avec les termes abstraits de notre langage. Qui peut se représenter concrètement ce qu'est la justice ? Une idée ? Ce qu'est l'énergie libre de Gibbs ? Un espace infini ? Il semblerait que notre connaissance abonde de mots sans représentation évidentes, sans qu'il soit besoin de parler de la conscience.

Ces contre-exemples ne sont toutefois pas en mesure d'ébranler ma position. Ni la justice, ni l'énergie libre de Gibbs, ni même l'infini ne sont des objets strictement paradoxaux. Ce sont certes des objets non représentables, mais non pas des objets contradictoires. C'est-à-dire qu'on peut les penser, on peut leur donner des propriétés logiques, on peut décrire les relations qu'ils entretiennent avec d'autres concepts. La conscience, en revanche, n'est pas pensable logiquement. Elle est concevable, mais pas pensable *stricto sensu*.

La Conscience face au Paradoxe de Russell

Dans les pages qui suivent, j'aimerais caractériser plus précisément le paradoxe de la conscience, de sorte à comprendre le lien qui existe entre les concepts de conscience vive, de conscience morte et les paradoxes logiques. Il est aisé de se représenter, à titre encore purement vague et intuitif, que le paradoxe de la conscience épouse la même structure que le paradoxe de Russell, aussi appelé le paradoxe du barbier (voir section 5.2.1). Ce dernier, rappelons-le, consiste à envisager l'ensemble R de tous les ensembles qui ne se contiennent pas eux-mêmes ; et à se demander si R se contient lui-même. S'il ne se contient pas, il doit se contenir. Et s'il se contient, il ne doit pas se contenir.

La structure du paradoxe de Russell semble être similaire au paradoxe de la conscience. La conscience vive est en effet définie comme l'ensemble de la réalité vécue. Or la conscience semble être quelque chose de vécu. Il faut donc que la conscience soit dans la conscience vive. Le paradoxe survient du fait que si la conscience est dans la conscience, elle se présente comme un simple objet de conscience. Ce n'est donc plus un sujet conscient. On en conclut que la conscience ne peut pas être dans la conscience ; mais qu'elle doit pourtant l'être. Si en effet la conscience n'est pas dans la conscience vive, alors elle n'existe pas comme vécu, ce qui semble contredire l'expérience d'être conscient.

Il est important de noter que ce paradoxe est plus profond qu'il ne peut le sembler à premier abord. Non seulement il affirme que la conscience n'est pas quelque chose de mental, puisqu'elle ne peut pas exister dans la conscience vive ; mais le paradoxe exclut également que nous puissions nous faire de la conscience une représentation scientifique matérialiste. Si en effet on détenait un modèle neurophysiologique de la conscience, on aurait une représentation explicite de la conscience. Cette représentation serait une image mentale de notre conscience. Or le paradoxe exclut justement que nous puissions nous faire une image objective de la conscience. Tout ce que nous pouvons dire, à la rigueur, si nous voulons vraiment contourner le paradoxe, c'est que la conscience existe de façon matérielle, ou autrement dit objective, dans un espace situé au-delà de la conscience — et que nous ne pouvons pas nous faire de cet espace une représentation explicite. Il s'ensuit, en des termes plus pragmatiques, l'une des alternatives suivantes :

1. La conscience existe comme une réalité matérielle, mais il n'existe pas de théorie matérialiste de la conscience.
2. La conscience existe comme une réalité matérielle, mais nous ne pouvons pas comprendre cette théorie.

Ces deux alternatives, comme le notera un lecteur consciencieux, coïncident avec les réponses que nous avons envisagées pour le dilemme de l'intuition dans le chapitre 9. Pour rappel, le dilemme de l'intuition consistait à se demander comment il est possible de donner à l'intuition une forme explicite qui échappe pourtant aux capacités de calcul d'une machine de Turing. Le problème consistait donc à expliquer, sous une forme explicite, pourquoi il y a quelque chose d'inexplicable dans l'intuition. Ce problème est fondamentalement le même que celui de la conscience, qui consiste à se demander comment on peut se faire une idée représentable de quelque chose de non représentable. Le difficile problème de la conscience et le dilemme de l'intuition ne sont que des manifestations d'un paradoxe logique plus fondamental.

La Conscience face au Paradoxe du Chat de Cheshire

Bien que l'on perçoive une analogie approximative entre le problème de la conscience et le paradoxe de Russell, j'aimerais donner un paradoxe logique qui épouse plus rigoureusement les propriétés spécifiques de la conscience. Ce paradoxe est de mon invention. J'aimerais vous l'introduire moyen d'un dialogue mettant en scène deux interlocuteurs. Pour rendre hommage à Charles Dodgson, qui a beaucoup contribué aux interactions entre mathématiques, philosophie et littérature, je prendrai le chat du Cheshire comme interrogateur et Alice au Pays de Merveilles comme répondante. Le paradoxe est obtenu au moyen d'un échange dialectique :

— CHAT : Jouons à une petite devinette. Je vais penser à un nombre naturel au hasard. Si tu le trouve, tu as gagné.

— ALICE : Mais comment veux-tu que je trouve ce chiffre ?

— CHAT : Essaie donc ! Je te dirai si ton chiffre est plus grand ou plus petit que le mien.

— ALICE : Très bien, dans ce cas, je dirais... voyons... 1000 peut-être ?

— CHAT : C'est un peu faible...

— ALICE : Un million ?

— CHAT : Plus encore...

— ALICE : Mais je chauffe au moins ?

— CHAT : Pas du tout !

Mon "paradoxe" stipule que le Chat ne cessera jamais de répondre "c'est un peu faible" quelle que soit la réponse d'Alice. Et pour cause : il existe une infinité de nombres entiers naturels. La proportion de nombres naturels inférieurs à la réponse d'Alice, aussi grande soit elle, est tout à fait négligeable par rapport à la proportion totale, et proprement infinie, des nombres naturels.

Il s'ensuit qu'à moins de donner à Alice (a) une infinité de tentatives (ce qui implique qu'elle et le Chat soient immortels) ou (b) d'assister à un miracle, il est impossible pour Alice de gagner le jeu. Ce "miracle" que j'évoque serait infiniment moins probable que d'obtenir l'intégrale des œuvres de Shakespeare en laissant un singe taper des lettres au hasard sur un clavier, comme Borel l'avait imaginé dans le paradoxe du singe savant en 1909. Ce serait en fait moins probable que tout événement ayant une probabilité, même infinitésimale, de se produire. Ce serait un événement mathématiquement possible, mais de probabilité strictement égale à zéro.

D'une certaine manière, il en va des nombres entiers comme de l'univers en expansion. Au fur et à mesure que l'espace-temps se dilate, les atomes qu'il contient s'éloignent les uns des autres. On trouve certes des amas de nébuleuses, mais après un temps suffisamment long, ce ne seront plus que des points isolés dans l'espace vide. Si l'expansion de l'univers progressait indéfiniment, la matière finirait par concentrer à sa surface. Et au fur et à mesure que cette superficie grandirait, la probabilité de trouver une particule à l'intérieur de l'Univers, et non à sa surface, tendrait vers zéro.

Bien sûr dans le cas de l'exemple qui nous intéresse, l'ensemble $\mathbb{N}$ des entiers n'est pas en expansion. Ce qui est en expansion, c'est l'énumération que l'on peut en faire en partant de 0 et en suivant une série de nombres croissants. On pourrait toutefois représenter sous forme d'une sphère l'ensemble des nombres inférieurs à la réponse d'Alice, qui définit le rayon de cette sphère. Tant que la sphère a un rayon fini, les nombres compris dans cette sphère sont en quantité négligeable. Mais dès qu'on donne à la sphère un rayon infini, elle comprend tous les nombres. La majorité de ces nombres, du reste, semble être à la surface de cette sphère ; ce que je veux dire par

là, c'est que la majorité des nombres entiers sont plus éloignés du centre que toute distance représentable. Il est dès lors paradoxal de comprendre où sont ces nombres exactement. Sont-ils à l'intérieur de la sphère ? Oui, ils le doivent, puisque ce sont des nombres finis. Mais ils ne peuvent pourtant pas être à aucune distance finie du centre. Ils sont donc à la fois dans la sphère et en dehors de la sphère.

J'aimerais défendre, dans les pages qui vont suivre, que le paradoxe de la conscience est exactement comme le nombre du Chat de Cheshire : la conscience objective est un objet auto-extensif ; un objet à la surface d'une sphère dont le rayon fuit vers l'horizon. Pour comprendre cette analogie, et contourner quelques objections techniques qu'on pourrait me faire en regard de la théorie des probabilités, il convient tout d'abord de poser quelques définitions rigoureuses.

Formalisation du paradoxe

Définition 28. Soit un ensemble $E \subset \mathbb{N}$ et $n \in \mathbb{N}$. La densité de E dans $\mathbb{N}$ se note δ_E et vaut :

$$\delta_E = \lim_{n \rightarrow +\infty} \frac{|\{e \in E \mid e < n\}|}{n}.$$

La densité δ_E définit une mesure de l'espace mesurable $(\mathbb{N}, \mathbb{P}(\mathbb{N}))$. Intuitivement, elle capture la probabilité, si on tire un naturel au hasard dans $\mathbb{N}$ (d'après une loi de probabilité uniforme), que ce naturel appartienne à E .

Exemple 8. Soit $A = \{2, 4, 8, \dots, 2n\}$ et $B = \{2, 4, 16, \dots, 2^n\}$. La densité de A dans $\mathbb{N}$ vaut $\frac{1}{2}$. Par contre, la densité de B dans $\mathbb{N}$ est nulle. Les deux ensembles ont la même cardinalité, mais en termes de densité, l'un est infiniment supérieur à l'autre. Topologiquement, on pourrait dire que B est un simple point de A .

Paradoxe 1. Soit un entier naturel m quelconque. Tirons au hasard un entier X dans $\mathbb{N}$ d'après une fonction de répartition uniforme $P(X < m) = \delta_{\{0,1,\dots,m-1\}}$. Il y a une probabilité égale à 1 que X soit strictement supérieur à m .

Démonstration. Tout d'abord, on remarque que la fonction de répartition est bien définie car $P(X \in \mathbb{N}) = 1$. Néanmoins, pour tout $m \in \mathbb{N}$, la probabilité que $X \in \{0, 1, \dots, m\}$ est donnée par :

$$\begin{aligned} \delta_{\{0,1,\dots,m\}} &= \lim_{n \rightarrow +\infty} \frac{|\{e \in \{0, 1, \dots, m\} \mid e < n\}|}{n} \\ &= \lim_{n \rightarrow +\infty} \frac{m+1}{n} \quad (\text{dès que } n > m) \\ &= 0. \end{aligned}$$

Il y a donc une probabilité égale à 1 que $X \in \mathbb{N} \setminus \{0, 1, \dots, m\}$, ou autrement dit que $X > m$. $\square$

Ce "paradoxe" est contre-intuitif dans la mesure où on s'attendrait à ce que, pour un certain entier naturel m suffisamment grand, il soit probable, ou du moins pas improbable, que $X < m$. Mais penser ainsi, c'est oublier que le nombre d'entiers naturels inférieurs à m est négligeable face au nombre d'entiers naturels supérieurs à m . Cela est d'ailleurs vrai que l'on compte en terme de densité, de cardinalité ou d'ordinalité. (voir l'annexe 4 pour une définition précise de ces trois notions).

Notre esprit fini nous empêche malheureusement d'imaginer des nombres un tant soit peu élevés, si bien que, lorsqu'on nous demande de prendre un naturel au hasard, nous le choisissons au sein d'un ensemble très restreint de possibilités. Or il va de soi que si nous le prenions *réellement* au hasard, ce nombre, et sans aucune contrainte sur la taille, il serait très grand. Tellement grand qu'il serait probabilistiquement impossible de donner un nombre supérieur.

Les Objets Auto-extensifs

On pourrait nous objecter que le Chat de Cheshire ne peut pas choisir un nombre naturel "au hasard", que cette expression n'a pas de sens. En mathématiques, il faudrait en effet définir une fonction de probabilité stipulant quelle sont ses chances de choisir tel nombre plutôt que tel autre. Or la fonction de probabilité ne peut pas être équiprobable pour un ensemble infini. Sa définition l'en empêche.

Cette objection est tout à fait recevable, mais je pense qu'elle ne nous interdit pas de penser, comme simple expérience conceptuelle, que le chat puisse utiliser une fonction de probabilité équiprobable. Il n'y a en effet rien de paradoxal, en soi, à imaginer une telle fonction. On arrive seulement à la conclusion que le nombre qui résulterait de ce choix aléatoires aurait la propriété particulière d'être bel et bien un nombre naturel — mais ce serait un naturel plus grand que tout autre. Il n'aurait donc pas de valeur précise.

Un tel nombre serait ce que j'appelle un objet *auto-extensif*. Chaque fois qu'on essaierait de connaître sa valeur précise, en posant une question à son égard (te trouves-tu dans tel intervalle?), on obtiendrait certes de l'information supplémentaire sur ce nombre. Mais cette information ne nous permettrait jamais de placer ce nombre dans un intervalle de taille finie.

On pourrait par exemple avoir une réponse à la question suivante : es-tu un nombre pair ? es-tu un multiple de trois ? admets-tu une racine carrée entière ? Ces questions recevraient une réponse claire et précise. Mais il resterait toujours une infinité de possibilités. Autrement dit, il n'existe aucune combinaison de questions qui permettraient de situer ce nombre auto-extensif dans un intervalle fini.

À cet égard, les nombres auto-extensifs se comportent comme la conscience objective. On peut poser des questions à leur égard, et recevoir des réponses, mais sans jamais dissiper toutes les inconnues demeurant à leur endroit. On peut par exemple dire qu'il est nécessaire que le nombre du Chat de Cheshire soit plus grand que tel entier n donné ; ou qu'il soit un multiple de trois. Mais on ne peut pas trouver de conditions à la fois nécessaires et suffisantes. Il s'ensuit que ces objets auto-extensifs répondent, par nature, à la définition du mystérianisme de Jesse Prinz que j'ai discuté dans la section 9.4.4, et qu'il récapitule lui-même avec beaucoup de clarté¹³ :

On peut savoir que boire deux litres d'alcool suffit à provoquer l'ivresse, et qu'au moins une cuillère à café d'alcool est nécessaire, sans pour autant connaître la quantité à la fois nécessaire et suffisante. Dans le cas de la conscience, les propriétés que nous savons suffisantes sont des propriétés dont les ordinateurs sont dépourvus. Les propriétés que nous savons nécessaires peuvent être possédées par les ordinateurs, mais nous ignorons si elles suffisent.

Ce détour par le paradoxe du chat de Cheshire me permet de résumer l'hypothèse 2 sous la forme suivante :

Hypothèse 5. *La conscience objective, ainsi que tous les objets paradoxaux, sont des objets auto-extensifs. Chaque fois que l'on tente de les caractériser, ils dépassent de la représentation qu'on peut leur donner. Il est certes possible de collecter de l'information à leur endroit, mais cette information sera négligeable. D'un point de vue géométrique, ils occupent la frontière d'une sphère croissant infiniment.*

12.4 Hypothèse 3 : Monisme à double aspect

La dualité de la conscience, qui peut tantôt être envisagée comme conscience vive, tantôt comme conscience morte, me conduit à formuler une troisième hypothèse : celle selon laquelle il existerait bien une réalité tangible à l'origine des concepts dichotomiques tels que la « pensée » et la « matière », ou « l'objectivité » et la « subjectivité », mais que ces concepts ne parviennent pas à capturer cette réalité de façon à la fois complète et cohérente. Les raisons qui me poussent à penser en ce sens sont multiples.

Tout d'abord, on observe que chacun des deux registres que sont le « Je » et le « Il » ne peut nous renseigner que partiellement sur la question de la conscience, sans jamais donner le dernier mot. Ils nous informent tantôt (i) sur la vie intime du sujet et les liaisons existantes entre ses contenus de conscience, tantôt (ii) sur les éléments de matière et leurs relations de

13. traduit par mes soins à partir de l'article *Level-headed mysterianism and artificial experience* de Jesse Jesse Prinz.

causes à effets. Ils nous parlent en revanche beaucoup plus difficilement des rapports entre les deux aspects. On soupçonne qu'il doit bien exister un lien, ou une structure commune, qui explique les corrélations observées entre certaines activités cérébrales et la vie mentale auto-rapportée, mais dès lors que, partant de l'un de ces registres, on tente d'expliquer un phénomène à l'aide des concepts issus de l'autre registre,

Ce qu'on cherche à gagner en complétude se paie au prix d'une certaine incohérence dans les termes, rendue manifeste par la présence d'expressions linguistiques symptomatiques. Il en va ainsi, par exemple, des erreurs méréologiques observées dans les écrits de certains neuroscientifiques, lorsqu'il est dit du cerveau qu'il « décide » (Damasio¹⁴), qu'il « croit », « interprète », « devine » (Crick¹⁵), qu'il « connaît », « raisonne », « construit des hypothèses » (Blackmore¹⁶), qu'il « catégorise », « distingue » ou « manipule des règles » (Edelman¹⁷), ou lorsqu'il est affirmé, des neurones eux-mêmes, qu'ils sont « intelligents » ou qu'ils « estiment des probabilités » (Blackmore¹⁸). Bien qu'utiles à titre analogique, ces expressions trahissent souvent la présence d'un chiasme métaphysique externaliste, puisque des états intentionnels sont attribués directement aux manifestations extérieures, ou physiologiques, du cerveau. Ce genre d'erreur est également désignée, dans la littérature anglophone, sous le nom « mereological fallacy¹⁹ ».

Contrairement à ce que prétendent certains auteurs, nous ne pensons pas que ce soit *uniquement* à cause d'abus de langage que le chiasme se produise — et qu'il suffirait en quelque sorte d'épurer les théories neuroscientifiques de l'usage de certains termes pour résoudre le problème *ipso facto*²⁰. Nous souhaitons alléguer, à l'encontre de cette opinion, deux justifications.

La première, c'est que si l'analogie qui est établie entre les faits mentaux et les données physiologiques était purement linguistique, les propositions au sujet de la conscience perdraient une grande part de leur signification, dans la mesure du moins où elles ne nous éclaireraient plus vraiment à propos des causes de l'existence de cette vie mentale dont nous disposons. Ce geste, qui consiste en fait à évacuer l'erreur du raisonnement en même temps que

14. DAMASIO A., *L'Erreur de Descartes*, Odile Jacob, 1997, p.69

15. CRICK F., *Astonishing Hypothesis : The Scientific Search for The Soul*, Cambridge University Press, Cambridge 1995, p.30.

16. BLACKMORE C., *Mechanics of The Mind*, Touchstones, Londres 1997, p.91.

17. EDELMAN G., *Bright Air, Brilliant Fire : on The Matter of the Mind*, Penguin, Harmondsworth 1997, pp.109

18. BLACKMORE C., *Ibidem*.

19. DANIEL BENNETT M., DENNET D., HACKER P., SEARLE J., *Neuroscience and Philosophy : Brain, Mind, and Language*, Columbia University Press, 2003.

20. Simon Ullman défend que les propriétés psychologiques imputées au cerveau relèvent seulement d'un certain vocabulaire technique (c.f. *Tacit Assumptions in the Computational Study of Vision*, p. 314.), Richard Gregory et Ernst Gombrich, pour leur part, défendent qu'il s'agit d'analogies (*The Confounded Eye*, p. 51.) et Blackmore parle de métaphores (*Understanding Images in the Brain*, pp. 257–83.)

le problème auquel ce raisonnement cherchait à répondre, nous laisse avec un arrière-goût d'insatisfaction, puisqu'il met entre parenthèses l'objet de la démarche du neuroscientifique, sa visée sincère et inavouée : celle de rendre compte du lien obscur existant entre le corps et l'esprit.

La deuxième raison pour laquelle il est difficile d'expliquer ces expressions frauduleuses en termes purement linguistiques, c'est qu'on s'expose au risque de s'appuyer maladroitement sur une forme de positivisme, ou encore, de supposer dangereusement qu'il existe comme une frontière arrêtée entre le monde des mots et le monde des choses. L'existence même d'un prétendu « abus de langage » signifie en effet qu'un certain mot est transposé de façon erronée à un domaine d'objets auquel il ne devrait pas se référer. Il faut donc, d'une part, que les domaines des objets mentaux et des objets physiques puissent être définis préalablement au langage ; et d'autre part, que le sens des propositions puisse être ramené, par un jeu de correspondance savamment construit, à ces objets signifiants.

Cela implique de définir premièrement le domaine des objets possibles, la sémantique, puis d'inférer le domaine des expressions linguistiques valides, c'est-à-dire une syntaxique. Or non seulement la distinction entre proposition synthétique et syntaxique n'est pas évidente à produire (c'était l'objet de la première partie de cette thèse), mais il semblerait en outre que ce soit par le jeu du langage que nous cherchons à les produire ; si bien que les mots conditionnent dans une certaine mesure la sémantique que nous cherchons à leur détacher. Au vu de ces difficultés, n'est-il pas raisonnable de prendre pour argent comptant les expressions linguistiques que le sens commun et la pratique neuroscientifique a mises à notre disposition et, plutôt que de chercher à les disqualifier, de nous poser simplement la question suivante : par quel jeu de circonstances ces expressions, d'apparence abusive, peuvent être dotées d'une signification ?

Il s'agirait dès lors non pas d'écarter, à la façon d'un inquisiteur, les propositions qui sont dénuées de sens, mais d'en chercher les conditions d'intelligibilité. Une telle démarche aurait pour conséquence de remonter aux concepts d'esprit et de matière, d'objectivité et de subjectivité, et de se demander s'il est possible de parler de la conscience morte sans parler dans le même temps de la conscience vive. Ou autrement dit, s'il est conceptuellement possible d'isoler l'une de l'autre, ou si au contraire ces concepts sont si intriqués que toute expression traitant de l'une contient, dans ses conditions même d'intelligibilité, une référence à l'autre ?

Pour donner à ce point de vue une illustration tangible, il peut être utile d'avoir à l'esprit le simple exemple des concepts « d'intérieur » et « d'extérieur ». Lorsque, dans un usage topologique de l'espèce la plus évidente, je formule une phrase du type « A est extérieur à B », le sens même de la phrase suppose qu'il soit possible de comprendre les conditions selon lesquelles A ne serait pas extérieur à B, c'est-à-dire, selon lesquelles A serait intérieur à B. Je ne peux ainsi définir ce qui est « extérieur » qu'en référence à ce qui

est « intérieur », si bien qu'il est impossible d'avoir un discours qui scinde complètement les deux concepts l'un de l'autre.

Bien sûr, la connaissance de ce qui est extérieur à B n'implique pas une connaissance complète, où même partielle, de ce qui lui est intérieur, mais uniquement de ce que ce serait pour A que d'être intérieur à B. Cette nuance est importante car elle revient à dire que ce ne sont pas les objets extérieurs à B qui sont intriqués aux objets intérieurs à B, mais seulement les concepts d'intérieur ou d'extérieur eux-mêmes, c'est-à-dire, en d'autres termes, les conditions de possibilité pour un objet d'être intérieur ou extérieur à B.

De façon tout à fait identique, la connaissance complète de tout ce qui peut être dit de vrai, ou pourrait être dit de vrai, dans une perspective en « Il », n'implique pas une connaissance complète, ou même partielle, de tout ce qui pourrait être dit de vrai dans une perspective en « Je » ; et il n'est même pas nécessaire de supposer un lien de cause à effet entre les objets de ces deux perspectives. En revanche, comprendre qu'une chose est vraie dans la première perspective consisterait à comprendre ce qui pourrait rendre cette chose fausse.

Or s'il en va de la perspective en « Je » et la perspective en « Il » comme des concepts d'intérieur et d'extérieur dont nous avons parlés ci-dessus — d'une manière qu'il reste encore à élucider — cela signifie que comprendre ce qui rend vraie une proposition dans la perspective en « Il » (l'extérieur) nécessite de faire référence aux conditions de vérité de la perspective en « Je » (l'intérieur). Pour le dire autrement, notre hypothèse consisterait à supposer que les points de vue du « Il » et du « Je » sont intriqués conceptuellement l'un à l'autre, d'une manière telle que l'un ne peut être rendu intelligible qu'à la lumière du second.

S'il est facile de voir comment des concepts duaux tels que « l'intérieur » et « l'extérieur », le « froid » et le « chaud », ou le « grand » et le « petit », peuvent être intriqués conceptuellement, il semble en revanche plus complexe de le percevoir dans le cas de la conscience morte et de la conscience vive. Les premiers en effet semblent n'être rien d'autre qu'une formulation tantôt positive, tantôt négative, du même phénomène, exactement comme lorsque l'on parle d'un « verre à moitié vide » ou « à moitié plein²¹ ».

Or dans le cas de la conscience morte et de la conscience vive, l'intrication ne peut pas être pensée selon le rapport d'un concept et de son contraire, mais elle n'en reste pas moins plausible (ou concevable) si l'on parvient à montrer que le point de vue en « Il » convoque, à un niveau plus fondamental,

21. Une remarque d'Harald Atmanspacher produite sur le concept de complémentarité pourrait s'insérer ici : « Notons cependant que, dans le cadre d'une logique booléenne, la complémentarité d'une proposition diffère de sa négation. Des exemples pertinents sont le complémentaire de la croyance, à savoir le doute plutôt que la non croyance (James), ou le complémentaire de la bonté, à savoir l'équité plutôt que la malignité (Bohr). » (ATMANSPACHER HARALD, *Le monisme à double aspect selon Pauli et Jung*, 2014-1, N. 3, pp. 105-135

le point de vue en « Je ». Comment exactement ? C'est ce qu'il convient de se demander à présent au moyen de quatre analogies.

12.4.1 Première analogie : L'Espace et les corps étendus

Pour essayer de comprendre le rapport d'intrication existant entre deux concepts, l'histoire de la philosophie a mis à notre disposition un certain nombre d'exemples inspirants. Aussi peut-on se demander, pour commencer, s'il n'en va pas du point de vue en « Je » et en « Il » comme du rapport entre l'espace et les corps dont il est constitué ?

De prime abord, l'espace et les corps qui l'habitent sont deux concepts distincts l'un de l'autre ; et pourtant, il semble tout aussi difficile de se représenter un corps étendu en dehors de tout espace, qu'il n'est difficile d'appréhender un espace sans corps. Si l'on en croit Kant, en effet, tout phénomène extérieur convoque, à un niveau fondamental, une certaine intuition pure de l'espace, qui lui est nécessaire et logiquement préalable.

À rebours de ce premier résultat, une idée centrale de la physique contemporaine consiste au contraire à dire de l'espace que, loin d'être ce contenant éthéré vivant indifféremment de son contenu, il ne serait finalement qu'un produit de la matière²². Ce renversement est intéressant, car il suggère que, d'un point de vue *a priori*, il est possible de considérer l'espace comme condition nécessaire de la connaissance sensible des corps ; tandis que d'un point de vue empirique, ce sont les corps au contraire qui définissent l'espace.

S'il en va bien ainsi, on tiendrait ici une intrication conceptuelle beaucoup plus proche de celle que nous recherchons, et la question reviendrait alors à demander si ce que nous avons appelé la conscience morte, et qui n'est rien d'autre que la conscience envisagée d'un point de vue extérieur, peut être conçue indépendamment de la possibilité d'une conscience phénoménologique.

Il convient d'explorer cette analogie et de voir si elle peut nous enseigner quelque chose sur la nature même des concepts. Cela demande, dans un premier temps, de montrer en quoi la perspective en « Je » peut-être conçue

22. Cette proposition peut être comprise en deux sens différents, selon que l'on se place du point de vue de la théorie de la relativité ou du point de vue de la mécanique quantique. On pourrait peut-être dire, avec Michel Paty, que « L'espace de la théorie relativiste de la gravitation (relativité générale) étant structuré par les champs engendrés par les masses des corps, n'était plus pensable en l'absence de corps physiques. On pourrait dire, en quelque sorte, que l'espace vide est constitué par les propriétés des corps en dehors du lieu de leur localisation, qu'il est le lieu matériel de l'influence des corps sans les corps. Quant à la physique quantique, elle en fait un substrat matériel à l'état latent qui peut se manifester en chacun de ses points sous une forme physique tangible (particules et champs) dès qu'un certain degré d'excitation est appliqué en ce point (polarisation du vide et création de paires) » (MICHEL PATY, *Le vide matériel, ou la matière crée l'espace*, Simon Diner et Edgard Gunzig, Univers du tout et du rien, Éditions de l'Université Libre de Bruxelles/Complexe, p. 22-44, 1998.)

comme une condition nécessaire de la perspective en « Il », puis de montrer la relation en sens inverse.

Étape déductive 1 : Le concept de conscience vive est condition de possibilité du concept de conscience morte

Prenons cette idée précautionneusement et raisonnons par étapes. Envisageons le cas d'un scientifique qui, prenant le problème de la conscience du point de vue naturaliste, entreprendrait l'expérience suivante : un certain sujet A énonce quel est son état de conscience E (ses peurs, ses idées, ses émotions ou représentations...) lorsqu'il est placé face à un certain stimuli S. Dans le même temps, l'expérimentateur prend note des mesures M enregistrées par un encéphalogramme.

L'enjeu est d'identifier des relations de cause à effet, ou tout au moins des corrélations significatives, entre les mesures M et les états E, tout en s'assurant que ces états soient stables en présence répétée du stimuli S. À titre d'exemple, les stimuli S pourraient consister en une injonction (pensez à votre sport préféré) et les mesures M pourraient être l'activité observée à l'aide d'un encéphalogramme dans le cortex moteur et somesthésique. Les états E pour leur part consistent à demander au patient quelles pensées ou quelles émotions ont accompagné l'injonction.

La question que nous adressons consiste à savoir si le dispositif expérimental qui vient d'être décrit pourrait remplir ses objectifs si l'expérimentateur ne disposait pas de l'idée d'arrière-plan, analogique, selon laquelle les états E rapportés par le sujet A sont similaires aux états que lui-même serait *susceptible* d'éprouver ? Je dis bien « susceptible » car la question n'est pas ici de savoir si l'expérimentateur éprouve effectivement la même chose que le patient A, ni de savoir s'il rapporterait le même registre de sensations E face au stimuli S. La question consiste à se demander si l'expérience aurait encore du sens en l'absence de la possibilité conjointe d'un point de vue à la première personne.

Pour tenter de répondre à cette interrogation, et en préciser du même coup les délimitations, supposons que le même dispositif expérimental dont nous parlions soit mis au point par un sujet incapable de toute intentionnalité — cette sorte de créature qu'on aime à appeler sous le nom de zombie philosophique, ou alternativement, par une machine dotée de capteurs sophistiqués. Il serait, dans ce cas, difficilement concevable que le zombie ou la machine acquiert, à propos de l'expérience qu'il est en train de mener, la même compréhension que celle acquise par un sujet intentionnel. Et pour cause : les signaux qu'il produirait et observerait pourraient être décrits par des règles formelles, ou prédits par des modèles, mais ils n'auraient pas à ses yeux plus de signification que les signaux émis par un électron.

Un tel zombie serait en fait réduit à un état comparable à celui décrit dans la chambre chinoise de Searle. Tout ne serait, pour lui, que manipula-

tions de symboles. Ces symboles renverraient certes à des objets du monde, mais non à des états mentaux qu'il pourrait expérimenter. L'idée même selon laquelle ces signaux seraient la manifestation d'une vie intérieure ne pourrait émerger que de façon strictement abstraite et intellectuelle dans son esprit — si tant est qu'il ait un esprit à proprement parler. Elle apparaîtrait un peu à la façon dont se présente l'idée, pour nous autres êtres humains, de ce que devrait être la vie d'une plante. On pourrait imaginer que la plante ressent quelque chose, mais ce terme « sentir » n'aurait de signification qu'en référence à notre propre expérience sensorielle, sans doute fort éloignée des végétaux.

Notre idée ici ne consiste pas simplement à dire qu'une expérience, en ce qu'elle convoque toujours, à un certains stade, une interprétation, ne peut pas être totalement dissociée des pensées et de l'état d'esprit de l'expérimentateur. Une thèse de cet acabit est généralement perçue comme un lieu commun et ne nous ferait, par ailleurs, pas significativement avancer dans notre propos. Ce que nous défendons, plus fondamentalement, c'est l'idée que les expériences que nous avons décrites ici, et qui porte sur la conscience, induisent une forme d'interprétation différente des autres expériences réalisées par les sciences naturelles.

Pour bien clarifier le propos, et exhiber du même coup les spécificités propres à l'étude empirique de la conscience, il convient de considérer, à titre d'archétype, une expérience de physique classique. Lorsque Galilée utilise les plans inclinés pour étudier la chute des corps, il ne cherche rien d'autre que d'établir des rapports entre différentes mesures, à savoir le temps qui s'écoule entre le moment où le mobile, partant d'une position initiale donnée, arrive à la fin de sa trajectoire, en fonction de l'inclinaison de la pente. Cette expérience ne requière pas de se référer aux états mentaux subjectifs d'un sujet. S'il est vrai que le sens de l'expérience dépend, dans une certaine mesure, des représentations que le scientifique se fait du mouvement, du temps ou de l'espace, les termes théoriques, en eux-mêmes, ne se réfèrent pas à ses états mentaux.

Dans le cas du dispositif neuroscientifique que nous avons décrit, en revanche, il y a présence d'une *théorie de l'esprit*²³. L'expérimentateur ne s'intéresse pas seulement à la conscience du patient comme à une boîte noire, dans laquelle certaines informations sont insérées en input (le stimuli S) et certaines autres ressortent en output (les mesures M et les états auto-rapportés E). En réalité, il cherche à inférer quelque chose sur son vécu « intérieur ». Pour comprendre le sens de ses conclusions, il est donc indispensable d'établir une analogie entre ce que je suis potentiellement capable de ressentir à la première personne, en tant que sujet, et ce que je pense que

23. Au sens de David Premack et Guy Woodruff lorsqu'ils écrivent : « Un individu a une théorie de l'esprit s'il impute un état mental à lui-même ou à autrui. » (voir l'introduction de leur article *Does the chimpanzee have a theory of mind ?*, Behavioral and Brain Sciences, Vol. 1, N. 4, décembre 1978, pp. 515-526). La traduction est de moi.

la patient de l'expérience prétend ressentir. Cette différence est précisément ce qui fait sortir les neurosciences d'aujourd'hui du paradigme behaviouriste qui régnait du temps d'Alan Turing et de l'invention du test portant son nom.

En conclusion de ce raisonnement, nous pouvons dire que le point de vue en « Il » qui est convoqué dans un modèle neuroscientifique portant sur la conscience semble n'acquiescer toute sa signification qu'en raison de la possibilité préalable du point de vue en « Je ». Derrière les représentations objectives à la troisième personne se tient toujours l'idée nécessaire d'une vie mentale possible.

Étape déductive 2 : Le concept de conscience morte est condition de possibilité du concept de conscience vive

Prenons maintenant la relation dans le sens inverse. En complément de ce que nous venons de dire, il se pourrait bien que le point de vue à la première personne ne puisse pas davantage être rendu intelligible sans la coexistence d'un point de vue à la troisième personne. Est-il possible en effet de recourir à l'introversion sans se référer indirectement à des concepts qui excèdent l'introspection ? L'introversion pure est-elle seulement envisageable ? Ou en d'autres termes, peut-on avoir un accès direct à notre propre subjectivité ?

Assez étrangement, cette critique semble plus facile à produire que celle qui concernant le paragraphe précédent. Peut-être en est-il ainsi à cause de l'histoire récente de la psychologie, et du sort peu enviable qui a été réservé à la méthodologie de Wundt²⁴ ? Peut-être pourrait-on dire, à la manière de Kant, que si tout psychologie commence avec l'introspection, d'un point de vue chronologique, elle n'en découle pas pour autant d'un point de vue logique ?

Une autre façon de voir le problème consisterait à prendre au sérieux la possibilité de se connaître comme substance pensante et de voir à quelles impasses cela nous conduirait. Il y a dans l'idée du cogito cartésien un certain paradoxe, relevé par nombre de commentateurs²⁵, qui consiste en cela que

24. Comme le dit très justement ARNAUD DEWALQ : « L'idée de Wundt est la suivante : en vertu des connexions causales existant entre le psychique et le physiologique, l'étude des processus physiologiques se trouve régulièrement renvoyée à des données de l'observation interne, donc à des données psychologiques. » (c.f. *Deux théories de l'analyse psychique : Wundt et Brentano*, In *Vers une philosophie scientifique : Le programme de Brentano* [en ligne]. Paris : Demopolis, 2014)

25. Kant le premier critique les conclusions de Descartes, et celles de la psychologie rationnelle en général, dans son chapitre concernant les paralogismes de la raison. Il écrit à propos de l'idée du cogito qu'elle est une idée de la raison pure sans contrepartie réelle : « Pour excellente que soit la conscience que nous avons de notre représentation de ces choses, il s'en faut encore de beaucoup que, si la représentation existe, l'objet qui lui correspond existe aussi... » (*Critique de la raison pure*, p. 382). De cette critique il découle que, chez Kant, il n'existe pas d'intuition intellectuelle. Seules les objets tombant sous le régime de la sensibilité pure disposent d'une forme *a priori* : l'espace et le temps. Pour

si on pense le vécu phénoménal comme la propriété d'un certain sujet, on se représente aussitôt ce sujet comme un objet dans le monde. La substance pensante est ainsi envisagée dans une spatialité, ce qui semble contraire à son essence même.

Ou pour le dire autrement, chaque fois que je cherche à me représenter ce que je suis en tant que pure intentionnalité, je convoque incidemment l'idée de moi-même comme étant quelque chose de substantielle. Cela ne veut pas dire que je suis effectivement un objet de l'espace, ni que l'espace existe en dehors de la pensée qu'on en a, car ce genre de questions est d'un autre ordre ; mais en revanche, cela induit que le concept d'objet, ou d'être soi-même un objet, est une condition de possibilité au concept de substance pensante. Pourrait-on en effet donner sens à la propriété d'être quelque chose de pensant si ce quelque chose ne pouvait pas être envisagé comme un objet ayant la propriété de penser ?

Il y a dans le cogito présence d'une sorte de paradoxe : d'une part le « Je pense » est le tout de la pensée, mais d'autre part il se pense lui-même comme un objet de la pensée. C'est comme s'il était à la fois le contenant et le contenu, celui qui pense et ce qui est pensé. Ce caractère bicéphale renvoie à la difficulté identifiée, chez Kant, de distinguer l'intuition formelle comme faculté subjective (espace métaphysique) de l'objet connu par l'exercice de cette faculté (espace géométrique). Tout se passe comme si le « Je pense » devait produire la preuve d'être autre chose que le contenu même de la pensée. Quoi exactement ? Un sujet sans doute, mais en quel sens ? Au sens d'être cette sorte de méta-propriété qui consiste, non pas à concevoir ceci ou cela de particulier, mais à concevoir des choses en général.

Or pour que le cogito puisse se penser comme une propriété, il doit se référer à lui-même comme à une objectivité. Cela implique d'abord de se différencier du monde qui l'entoure, et des idées qui le traversent, puis de reconnaître aux perceptions une sorte de permanence, et d'acquiescer enfin progressivement un concept d'objet. Mais comme il peut difficilement prendre de la distance par rapport à lui-même, et se regarder de l'extérieur, sans doute le sujet a-t-il besoin de recourir à toutes sortes de médiations ? À des représentations en quelque sorte étrangères à lui-même et sur lesquelles il peut appuyer son jugement ? C'est peut-être la raison pour laquelle des auteurs comme Husserl ou Mead, qui prennent au sérieux cette hypothèse d'une constitution de soi, insistent sur l'importance de l'intersubjectivité.

Ce phénomène, au reste, est observable même dans des situations assez simples de la vie ordinaire, lorsqu'on cherche par exemple se souvenir de ce que l'on pensait le jour précédent, et que tout à coup nous nous voyons là, au milieu de la pièce, comme au travers des yeux d'une tierce personne,

plus de développement, se référer notamment à AUCHATRAIRE ANNE, *Le cogito cartésien et la Critique de la raison pure*, in *Figures du Transcendantale*, dir. CRAMPE-CASNABET, MICHÈLE, Lyon : ENS Éditions, 1996, pp. 41-63.

accoudé à notre bureau ou à quelque chaise branlante, en train de penser à cette chose dont nous voulons nous souvenir. Le contenu même de notre pensée (ex : la pensée d'un chat) semble souvent être associé à la mémoire de l'endroit où nous nous tenions (je pensais à ce chat lorsque j'étais sur la chaise de mon bureau), et de la position que nous occupions (j'étais assis), au moment de le penser.

Il s'agit bien sûr que d'une métaphore sensible, et non d'un raisonnement apodictique. Mais il suggère que lorsque nous cherchons à avoir accès notre propre intériorité, à ce monde vivace de songes, de perceptions et de souvenir, nous nous référons aussi souvent, *dans le même temps*, au regard *possible* d'un autre individu posé sur nous. C'est-à-dire que nous nous voyons nous-même tel qu'un tiers pourrait nous voir.

À bien considérer la chose, il semblerait dès lors difficile de se penser comme le sujet des pensées, des sensations et des perceptions qui nous traversent, c'est-à-dire comme l'origine de tout contenu conscient, sans se voir dans le même temps de façon objective et extérieure à soi. Ou pour reprendre notre métaphore affectionnée, peut-être pourrait-on dire que penser l'intentionnalité en dehors d'un espace d'objectivité est aussi difficile que de penser un corps étendu en dehors d'un espace étendu *possible*.

Les deux concepts sont donc indissociables conceptuellement

Ce petit détour de la pensée, qui nous a conduit après du paradoxe du cogito et des zombies philosophiques, n'avait d'autre but que celui-ci : puisqu'on ne peut pas aisément penser le « Je » sans convoquer une représentation objective ; et puisqu'inversement on ne peut pas décrire la conscience dans une théorie en « Il » sans convoquer, au sein même de cette théorie, une analogie à notre propre vie mentale, il semblerait que les concepts d'objectivité et de subjectivité ne soient pas tout à fait dissociables conceptuellement.

Plutôt que de chercher à les traiter comme deux domaines distincts, ou de réduire l'un à l'autre, je suggère, conformément à notre hypothèse, de les envisager comme les deux faces complémentaires, quoique paradoxales, d'une réalité préconceptuelle sous-jacente. Tout se passe comme s'il n'existait finalement qu'une seule espèce de conscience, une conscience mort-vivante ; mais que celle-ci ne pouvait être pensée que sous une double modalité, celle de sujet ou celle d'objet. Cette conscience ne serait pas matérielle ou idéale à proprement parler, mais se présenterait davantage comme une structure intuitive, préconceptuelle, que nos formalismes linguistiques et scientifiques ne pourraient appréhender qu'au prix d'une certaine réduction méthodologique.

Notons que ce raisonnement que nous venons d'établir ne prouve en aucun cas la présence d'une réalité extérieure à nos sens, car il peut être reproduit de façon parfaitement cohérent au sein même de la régression phénoménologique transcendante qui est celle de Husserl ou de Descartes.

C'est d'ailleurs le résultat obtenu par Husserl lorsqu'il conclut, dans les méditations cartésiennes, que la conscience est toujours conscience *de quelque chose*.

Il serait du reste contradictoire, au vu des prémisses initiales de notre démarche, que de supposer ou de réfuter l'existence d'un « monde extérieur ». Vouloir montrer qu'il existe quelque chose au-delà de nos sens, d'un point de vue ontologique, c'est prendre le risque de sombrer dans une forme de dualisme, puisque pour avoir quelque chose en dehors de la pensée, il faut préalablement qu'il y ait quelque chose d'autre que de la pensée.

C'est pourquoi nous préférons dire que, d'un certain point de vue, tout est dans la pensée, mais qu'il y a, dans cette pensée, une idée qui réfère à quelque chose dont elle cherche à se séparer. Nous disons « d'un certain point de vue » et non pas « absolument », car il existe encore un autre point de vue selon lequel, au contraire, tout est matière et la conscience phénoménale n'existe pas. Ce point de vue, qui est celui de l'éliminativisme, n'est pas non plus contesté, *stricto sensu*, par notre argument, mais il est remis en perspective : l'idée selon laquelle la pensée émergerait de la matière ne prend-t-il pas appui sur un vécu de conscience préalable ?

Cette mise en perspective qui est opérée revient à demander, à celui qui nous explique que les contenus phénoménaux sont des illusions, d'où lui vient cette idée consciente qu'il nourrit à propos des contenus phénoménaux : cette idée n'est-elle pas déjà un contenu phénoménal ? Notre thèse n'a pas pour vocation d'invalider le point de vue de l'éliminativisme en ce qui concerne ses conditions de *vérité*, mais d'insister sur la présence, dans les conditions mêmes *d'intelligibilité* de la proposition comme quoi « les vécus phénoménaux sont illusoires », d'une référence nécessaire à quelque chose qui est phénoménal.

Notons encore, avant de finir, la présence d'un dernier parallélisme entre le concept d'espace et la perspective en « Je ». Du point de vue *a priori* qui était celui de Kant, il est en quelque sorte plus aisé de se représenter l'espace sans corps étendu, que des corps sans espace, puisque l'espace est ce sans quoi le corps ne peut pas être conçu. Tandis que du point de vue empirique, en revanche, le concept de corps semble premier par rapport à celui d'espace. Des mathématiciens comme Lobatchevsky ou Poincaré iront même jusqu'à dire que notre concept d'espace est déterminé par les expériences que nous avons, et dépend donc premièrement des corps²⁶. Or en ce qui concerne le

26. Penser par exemple à la conception empirique de l'espace défendue par Lobatchevsky, où l'espace n'est finalement qu'une construction mentale obtenue par la juxtaposition imaginée, et infinie, de plusieurs corps les uns à côtés des autres : « Lorsqu'ils se touchent, deux corps A et B forment un seul corps géométrique C, où les composants A et B restent distincts et ne sont pas perdues dans le tout de C. Inversement, une *section* arbitraire S divise un corps C en deux parties A et B... En ce sens, on peut imaginer tous les corps de la nature comme étant les parties d'un tout que nous nommons *l'espace*. » (LOBATCHEVSKY, *Nouveaux principes de géométrie avec une théorie complète des parallèles*, Kazan, 1835-1838.).

problème corps-esprit, il semblerait que, dans les approches idéalistes, il soit facile à accepter l'existence du cogito que des objets extérieures, tandis que d'un point de vue strictement empirique, on acceptera plus aisément l'idée des objets extérieurs que l'existence de sa propre conscience.

Le rapport existant entre la perspective en « Je » et la perspective en « Il » n'est toutefois pas, comme dans le cas du dualisme cartésien, du même ordre que la *res cogitan* et la *res extensa*. Au lieu d'être de natures différentes, et de se tenir imperméablement l'un en face l'autre, il en va de leurs rapports comme du Ying et du Yang. La démarche en « Je » privilégie le phénoménologique, mais tout en s'appuyant sur le naturalisme. De la même manière, la démarche en « Il » part du naturalisme mais prend forme grâce à la conscience phénoménale. C'est cette hybridation inhérente qui représente à la fois la meilleure objection au monisme physicaliste et la meilleure façon d'échapper à un dualisme sans issue.

12.4.2 Deuxième analogie : Les Théorèmes d'incomplétude

À ce stade, il pourrait sembler au lecteur que notre troisième hypothèse soit devenue difficile à soutenir ? Comment penser en effet un monisme dont les deux modalités seraient exclusives l'une de l'autre, tout en étant nécessaires à la compréhension du tout ? Aussi étrange qu'elle puisse paraître, cette idée tire pourtant son origine d'une intuition concernant la nature de la logique elle-même, et plus particulièrement des résultats d'incomplétude de Gödel, de Turing et de Tarski.

Certains des résultats les plus récents en logique des prédicats montrent que dans une théorie satisfaisant les théorèmes de Gödel, la probabilité, pour une proposition comportant n symboles, d'être indécidable, tend asymptotiquement vers le nombre un lorsque n tend vers l'infini²⁷. À partir de ce résultat, et sachant qu'il n'existe pas de borne à la taille d'une proposition bien fondée, il a été montré que, au sein de tout système formel suffisamment riche pour axiomatiser l'arithmétique, il existe infiniment plus de propositions indériverables que dérivables. En effet, la probabilité qu'une formule vraie soit indériverable en logique des propositions tend vers 1 lorsque la taille de la formule, en nombre de symboles, tend vers l'infini. De plus, l'ensemble des propositions démontrables, par rapport à l'ensemble des propositions indémontrables, est topologiquement du même ordre qu'un nuage de points sur une surface plane.

Bien sûr, une proposition qui est indémontrable au sein d'un système formel S peut très bien se démontrer dans un autre système formel S' . L'indécidabilité dont il est question ici n'est pas intrinsèque, mais relève strictement de l'axiomatique envisagée. C'est la raison pour laquelle il est efficace, pragmatiquement, d'admettre un pluralisme des théories. Dans la mesure où

27. C.f. CALUDE, JÜRGENSTEIN, *Is complexity a source of incompleteness ?*, Appl. Math. Comput. 66, pp. 63-76, 1994.

une même théorie est susceptible de rencontrer, en principe, beaucoup plus de propositions qui lui résistent que de propositions décidables, il semble raisonnable de croire (1) qu'il n'existera pas un seul système formel capable de rendre compte de la vérité de toutes les propositions qui en valent la peine ; (2) qu'un même problème peut être indécidable en regard d'un certain système formel, mais non en regard d'un autre.

L'évolution contemporaine des mathématiques semble d'ailleurs pointer du doigt les insuffisances inhérentes au fait de vouloir fonder la discipline sur une seule théorie axiomatique. Si la théorie de Zermelo-Fraenkel a gagné en popularité, il reste difficile de choisir les axiomes de façon à prouver tous les résultats intéressants, et uniquement ceux-là.

De la même façon, on observe en physique théorique une certaine difficulté à unifier la théorie de la relativité et de la physique quantique. Il se peut bien sûr que la physique ne réponde pas aux théorèmes d'incomplétude de la logique du premier ordre, ou que le problème de l'unification des théories ne relève pas de l'existence de propositions indécidables, mais la coexistence de plusieurs éclairages méthodologiques et disciplinaires sur une même problématique est devenue une pratique courante, pour ne pas dire souhaitable, dans de nombreux champs d'investigations — comme en atteste l'existence même des sciences cognitives, qui allient philosophie, linguistique, psychologie, neurosciences, anthropologie et informatique.

Il n'y aurait donc pas plus de contradiction à concevoir l'existence d'un monisme à double aspect qu'il n'y en aurait à concevoir l'édifice entier des mathématiques. Ce ne serait finalement rien d'autre qu'une structure composée d'une grande variété de contenus intuitifs plus ou moins formalisés, que l'on pourrait expliquer au moyen de deux systèmes axiomatiques concurrents. L'un partirait de l'axiome du *cogito* et l'autre partirait de l'axiome que tout est matière. Néanmoins, ils parleraient bien des mêmes objets.

D'ailleurs, les résultats de ces deux systèmes coïncideraient sur un vaste ensemble de questions, à vrai dire la majorité des questions constituant le savoir humain actuel, mais buteraient aussi sur un nombre infini de problèmes indécidables. Ces derniers concerneraient, pour l'essentiel, des métapropriétés, comme celle de savoir si la perspective en « Il » est capable de donner une réponse à tous les problèmes qu'il est possible de formuler dans cette perspective — ce qui est, par essence, le problème du réductionnisme des sciences de la nature.

12.4.3 Troisième analogie : Le Principe de complémentarité

Cette relation spécifique que nous tentons de décrire entre le point de vue en « Je » et le point de vue en « Il » peut encore recevoir un autre éclairage, inspiré cette fois de la physique quantique. Dans un article de 1952, le physicien Wolfgang Pauli, connu pour avoir établi le principe d'exclusion,

compare le problème corps-esprit au problème onde-particule^[28]. Plutôt de penser l'esprit comme résultant de la matière, ce qui fait intervenir une causalité croisée, ou de supposer à la manière de Leibniz une harmonie entre l'ordre de la pensée et de l'étendue, il préfère recourir à une extrapolation du principe de complémentarité de Bohr^[29]:

Le problème général de la relation entre la psyché et la matière, entre l'intérieur et l'extérieur, ne peut guère être considéré comme résolu par le terme de « parallélisme psychophysique » avancé au siècle dernier. Pourtant, peut-être, la science moderne nous a-t-elle rapprochés d'une conception plus satisfaisante de cette relation, car elle a élaboré la notion de complémentarité au sein de la physique. Il serait des plus satisfaisants que matière et psyché puissent être conçues comme des aspects complémentaires d'une même réalité.

La complémentarité qui est visée ici n'est pas liée à la non-commutativité des opérateurs quantiques, qui sont à l'origine de la violation des inégalités de Bell ou des états de superposition de la fonction d'onde. Il s'agit bien plutôt de la complémentarité onde-particule, selon laquelle un même phénomène admet deux descriptions exclusives l'une de l'autre, mais conjointement nécessaires. Ces deux descriptions sont requises pour obtenir une compréhension exhaustive du phénomène, mais sans être simultanément accessibles. Selon Velmans, c'est dans ce sens-là que le vécu expérimentiel d'une certaine personne et les activités cérébrales de cette personne, observées par un tiers, peuvent être dites complémentaires^[30].

Dans une abondante correspondance avec Carl Jung, Pauli va tenter de systématiser cette approche en s'appuyant sur l'idée que la mesure, en physique quantique, opère un découpage précis, mais imparfait, du phénomène considéré dans sa totalité — privant en cela les objets observés de leur plus haute connectivité. Son intuition fondamentale peut se formuler, dans la théorie quantique algébrique, à l'aide des états ontiques et épistémiques.

Les états ontiques caractérisent le système tel qu'il est dans sa globalité, mais ils ne sont pas observables en tant que tels. Les états épistémiques en revanche se rapportent à un contexte empirique particulier, et peuvent ainsi faire l'objet d'une observation au travers d'un protocole opérationnel. Le problème vient de ce que le contexte empirique procède à une rupture de la haute connectivité des états ontique. Des objets se dessinent à l'issue de la mesure qui ne capturent plus forcément toutes les facettes de l'objet

28. JUNG CARL GUSTAV, PAULI WOLFGANG, *Natureerklärung und Psyche*, Zurich Rascher, 1952. Pour une discussion philosophique plus approfondie de leur article, se référer à Atmanspacher Harald, *Le monisme à double aspect selon Pauli et Jung*, 2014, N. 3, pp. 105-135

29. *Ibidem*, p. 164

30. VELMANS MAX, *Understanding Consciousness*, second edition, Routledge/Psychology Press, 2009.

étudié. Cela induit l'idée qu'il existerait un réel dont la nature ne serait pas proprement mentale ou physique, mais qui se laisserait observée par l'entremise de ces deux points de vue.

Pauli propose ainsi d'établir une analogie avec la façon dont le réel se donne à voir, ou disons plutôt « à mesurer », dans la physique quantique, et le rôle occupé par l'inconscient dans les théories Jungiennes et, plus généralement, dans la psychanalyse³¹ :

« [...] la situation de la théorie de la connaissance portant sur les concepts de 'conscience' et 'd'inconscient' paraît présenter une analogie très poussée avec la situation de "complémentarité" en physique esquissée ci-dessous. D'une part, assurément, l'inconscient ne se laisse déceler qu'indirectement grâce à ses effets (ordonnateurs) sur des contenus de la conscience, d'autre part, toute "observation de l'inconscient", c'est-à-dire toute opération consistant à rendre conscients des contenus inconscients exerce une réaction tout d'abord incontrôlable sur ces contenus inconscients eux-mêmes (ce qui, comme on le sait, exclut par principe un "tarissement" de l'inconscient par le fait de rendre conscients ses contenus). Le physicien conclura donc par analogie que cette réaction incontrôlable du sujet qui observe sur l'inconscient limite précisément le caractère subjectif de sa réalité et confère en même temps à celle-ci une certaine subjectivité. Bien qu'en outre la place de la "coupure" entre la conscience et l'inconscient (tout au moins jusqu'à un certain point) soit laissée au libre choix de l'expérimentateur en psychologie, l'existence de cette coupure demeure une nécessité inéluctable. D'après ces données, le système observé ne se composerait pas seulement, du point de vue de la psychologie, d'objets physiques, mais il comprendrait en outre l'inconscient, tandis que le rôle assigné à la conscience serait celui de "moyen d'observation".

Par la suite, il semblerait que la tradition physicienne et psychanalyste se soient dissociées. Toute une lignée de chercheurs va poursuivre les idées de Pauli en recherchant des variables cachées dans les formalismes de la mécanique quantique (Bohm) ou dans son algèbre (Basil Hiley, Pylkkänen). D'autres, comme le physicien Bernard d'Espagnat, vont s'en emparer pour adresser le problème corps-esprit. Mais peu d'auteurs à ma connaissance semble avoir cherché du côté des formalismes utilisés dans les neurosciences contemporaines.

S'il est vrai que certains neuroscientifiques se réclament favorables au monisme spinoziste³², qui représente historiquement la première forme de

31. CARL JUNG, *Naturerklärung und Psyche*, Zurich Rascher, extrait de *Atmanspacher Harald, Ibidem*.

32. ATLAN est un vigoureux partisan du spinozisme. Voir son ouvrage publié en 2018.

pensée la plus proche de celle suggérée par Pauli, cela ne constitue pas une vision philosophique systématisée et actualisée. Il n'est pas sûr que la philosophie des sciences contemporaines puisse aisément se permettre d'adhérer à l'essentialisme du spinozisme, ou à sa conception de la causalité inspirée de la physique newtonienne.

C'est pourquoi il semble intéressant, pour tester la viabilité de cette hypothèse que nous avons introduite, concernant le rapport entre la conscience vive et la conscience morte, de voir s'il est possible d'exhiber des symétries de structures entre les théories neuroscientifiques contemporaines et le vécu expérientiel à la première personne. Mais avant d'envisager cette piste, qui constituera la deuxième partie de la thèse, il convient de donner encore une quatrième et dernière analogie.

12.4.4 Quatrième analogie : L'Hypersphère de Dante

Une dernière analogie peut être obtenue en explorant plus sérieusement l'idée que la conscience puisse être observée depuis deux perspectives. Ne pourrait-on pas imaginer un objet si complexe à se représenter dans sa totalité que l'esprit humain est contraint d'en étudier plusieurs sections ? Pour donner une métaphore facile, supposons que nous soyons réduits à observer un cylindre en trois dimensions dans un espace à deux dimensions. Que verrait-on ? La projection du cylindre sur une feuille de papier donne tantôt l'image d'une sphère, tantôt celle d'un carré, selon l'orientation du cylindre par rapport au plan de la feuille. Nous appelons ce carré « matière » et nous nommons « esprit » le cercle — quand nous ne nions pas purement et simplement son existence. Il s'agit pourtant bel et bien du même objet.

Pour poursuivre la métaphore un cran plus loin, essayons maintenant d'imaginer quelles formes prendraient un objet à quatre dimensions, telle qu'une hypersphère, si on la projetait dans notre espace à trois dimensions ? La meilleure façon de se représenter l'hypersphère à moindres frais, et avec une pointe d'humour, est sans doute de jeter un coup d'œil dans les Enfers de Dante. Le poète italien, jamais avare de tours de passe-passe, représente les enfers à l'intérieur de la Terre et le monde céleste à l'extérieur de la Terre. La sphère des enfers est donc comprise dans la sphère du paradis aussi sûrement que la Terre est dans l'Univers. Si Dante part du centre de la Terre, où réside Satan, et qu'il remonte à la surface en ligne droite, où se trouve-t-il ? À la surface des Enfers et du Paradis. Ce point de contact entre les deux mondes coïncide, chez Dante, avec la cité de Jérusalem. Si l'on part d'un point quelconque du paradis et que l'on progresse linéairement dans la même direction, où débouche-t-on ? Dans un monde en trois dimensions, on ne sort jamais du paradis. Après tout, l'Univers de Dante n'est-il pas infini ? Mais dans le cas d'une hypersphère de la quatrième dimension, on atteint la surface des enfers.

Vous m'aurez compris : l'hypersphère est un objet infernal ! Elle doit être

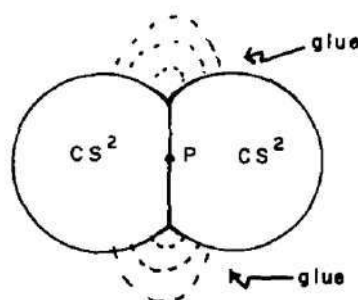


FIGURE 12.2 – L'hypersphère peut être vue comme deux sphères comprises l'une dans l'autre. La seule façon d'entrer dans l'une, c'est d'atteindre la limite extérieure de la seconde. En marchant toujours dans la même direction, on passe donc d'une sphère à l'autre de façon cyclique.

vue comme deux sphères connectées l'une à l'autre. Chaque fois que l'on sort de l'une, l'on pénètre dans la seconde. Par suite de quoi les concepts « d'intérieur » et « d'extérieur » changent de sens : la sphère des enfers est dans le paradis ; mais la sphère du paradis est elle-même dans les enfers, et ainsi de suite. Cette propriété inhabituelle n'est pas sans rappeler le rapport liant la conscience vive de la conscience morte. La conscience vive semble en effet « dans » la conscience morte, puisque les phénomènes se déroulent à l'intérieur de notre tête. Mais la conscience morte est elle aussi « dans » la conscience vive, puisqu'en effet le monde existe uniquement dans la conscience que nous nous en faisons.

Je soutiendrai que la conscience est exactement cela : un objet non représentable dans notre intuition en trois dimensions. Nous ne pouvons pas nous en faire une idée complète, mais seulement des projections métaphoriques. L'une de ces projections, c'est l'espace matériel. La seconde, c'est l'espace mental. Ces projections sont intriquées. Elles semblent contenues l'une dans l'autre, mais elles appartiennent en réalité au même objet. Je tenterai de définir une méthode hybride, la mystique rationnelle, capable d'entreprendre une reconstruction systématique, quoique laminaire et hypothétique, de l'unité sous-jacente aux deux dimensions.

Remarquons que j'ai utilisé le terme « dans » avec des guillemets. Cette précaution vient de ce que l'on entend souvent dire que l'esprit est « dans » le cerveau alors que je n'ai pas cessé, tout au long de ces pages, de suggérer l'inverse. Je laisse aux politiciens le soin de faire la girouette. Pour ma part, je continuerai de critiquer que ma tête puisse contenir une portion de conscience. Ce n'est ni un seau rempli d'une eau magique, ni une lampe renfermant un génie, mais l'objet d'un mystère autrement plus complexe : « Le fait que quelque chose d'aussi remarquable qu'un état de conscience, disait

Huxley, se produise par suite de l'initiation d'un tissu nerveux est tout aussi inexplicable que l'apparition du Djinn, après qu'Aladin eût frotté la lampe magique. »

Ce qu'exprime la métaphore de l'hypersphère, c'est que la totalité de l'espace psychique, prise comme un tout, est contenue dans la totalité de l'espace matériel, et inversement. Si bien que le scientifique ne trouvera jamais, en découpant les neurones, quoi que ce soit qui s'apparente à la vie consciente. Au contraire, pour trouver la présence du mental, il devrait atteindre les limites de la matière. Il peut avoir l'impression d'approcher cette limite en regardant, non pas les substances à l'intérieur du crâne, mais l'enveloppe du cerveau lui-même, pris comme un tout organique. Une fois engagée dans cette voie, il doit cependant, s'il veut être rigoureux, étendre le raisonnement à l'ensemble de la matière connue, en considérant tour à tour l'enveloppe du crâne, l'enveloppe du corps humain, puis l'enveloppe de la nature. Cette stratégie est celle qui fait dire à Clark que l'esprit n'est pas *seulement* dans le cerveau ou dans le corps, mais *aussi* dans l'interaction avec l'environnement :

*Il devient de plus en plus difficile de dire où le monde s'arrête et où commence l'individu*³³.

Cela peut donner l'impression à l'anatomiste en herbe que la conscience est à la fois partout et nulle part, immanente et transcendante, réelle et illusoire. Après tout quoi d'étonnant ? Rappelons-nous la métaphore de Dante. C'est en marchant toujours tout droit que le voyageur sortira de l'enfer des atomes pour pénétrer le paradis de l'esprit. Or il est tout aussi impossible de sortir de la matière en marchant dans l'espace qu'il n'est pas impossible de penser le monde matériel en dehors de la pensée.

Tout se passe donc comme si la vie consciente occupait les frontières infiniment éloignées de l'espace matériel et que la matière, en retour, occupait les frontières infiniment éloignées de la vie consciente. Plus exactement, je dirai que la la frontière intérieure de l'espace mental, qui est visible depuis cet espace, quoi qu'inatteignable, c'est la conscience vive. Tandis que la frontière extérieure du mental, qui est tout à fait invisible dans le mental, c'est la conscience morte. Cela explique que l'on ne trouve rien, dans notre conscience, qui s'apparente à un cerveau, à des nerfs, à des ganglions. Tout ce que nous voyons de ces objets inertes, ce sont des phénomènes de conscience : des idées, des perceptions, des représentations, etc.

La frontière visible de la matière, inversement, c'est la conscience morte. C'est celle qui se laisse apercevoir à la surface des corps, lorsque nous remarquons qu'ôter certaines parties du cerveau, ou soumettre l'organisme à certaines enzymes, semble avoir un effet sur le fonctionnement de l'esprit. La frontière invisible de la matière, en revanche, c'est la conscience vive. C'est le

33. ANDY CLARK, *Natural-born Cyborgs : Minds, Technologies, and the Future of Human Intelligence*, ed. Oxford University Press, USA, 2004. La traduction est de moi.

vécu. C'est ce qu'on arrive pas à localiser dans la matière. C'est cette âme invisible, intangible et immatérielle dont rêvait Swendenborg.

Il est vrai, dans l'espace à quatre dimension, que l'esprit est dans la matière et réciproquement. Mais dans notre réalité humaine, nous faisons bien plutôt l'expérience que l'esprit n'est ni vraiment « dans » la matière ni vraiment « hors » de celle-ci. Tout au plus lui est-il co-extensif. Le corps et le mental sont les deux facettes distinctes d'une réalité unique, mais irréprésentable. Tout au plus pouvons-nous en donner des paraboles, des évocations, des métaphores. Les notions « d'intérieur » et « d'extérieur », dans ce contexte, s'apparentent d'ailleurs plus au vocabulaire extasié de maître Eckhart, d'Ibn'Arabi ou des Upanishads qu'à notre expérience commune du monde. Et encore que ce vocabulaire puisse faire le même effet qu'un nuage qui obscurcit le soleil, je vais m'efforcer de montrer que les nuages ne gênent que ceux qui demeurent sur le sol de la pensée représentationnelle.

Il n'existe que deux façons d'expliquer comment un cerveau fini est capable d'exécuter les fonctions de l'intuition :

- Soit en modélisant le cerveau avec le formalisme des machines transfinies ;
- Soit en recourant à un oracle (une boîte noire).

Ces deux approches sont insatisfaisantes. Je me propose dès lors de défendre un *mystérianisme de perspective*. Celui-ci consiste à dire qu'on peut avoir un modèle du cerveau et un modèle de l'esprit, mais qu'il n'existe pas de modèle expliquant la liaison entre le corps et l'esprit.

Pour justifier ce point de vue philosophique, je fais trois hypothèses sur la conscience :

1. La conscience peut être conçue de deux manières : soit de façon *subjective* (conscience vive) soit de façon *objective* (conscience morte).
2. Il existe un paradoxe logique découlant de la définition de ces deux concepts. Ce paradoxe est lié au paradoxe de Russell.
3. La conscience doit dès lors être envisagée dans deux perspectives : une perspective en « Je » et une perspective en « Il ».

Dans le prochain chapitre, il s'agira de montrer que cette vision de la conscience est compatible avec l'évolution des neurosciences.

Chapitre 13

Critique de l'encéphalocentrisme

Non seulement nos plaisirs, nos joies et nos rires mais également nos tristesses, nos peines et nos chagrins proviennent du cerveau, et du cerveau uniquement.

Hippocrate

Dans le monde neuroscientifique, il arrive de faire l'amalgame entre les concepts de conscience, d'esprit et de cerveau. Le meilleur exemple de ce phénomène est peut-être à chercher dans les traductions francophones des ouvrages scientifiques ? « Bright Air, Bright Fire : on the Matter of the Mind » de Gerald Edelman est traduit « Biologie de la Conscience », tandis que « Wider Than the Sky. The Phenomenal Gift of Consciousness » et devenu « Plus vaste que le ciel. Une théorie générale du cerveau ». Ces glissements sémantiques trahissent un croisement hybride entre une métaphysique explicitement matérialiste et des références implicites à des expériences phénoménologiques à la première personne.

Dans la deuxième hypothèse sur la conscience (section [12.3](#)), j'ai défini le chiasme externaliste comme la propension à confondre une *description* purement fonctionnelle du cerveau, peuplé de myéline, de synapses, d'axones, de neurotransmetteurs et de potentiels électriques, avec une *explication* proprement convaincante du phénomène de la conscience. Tout se passe, au sein de ce chiasme, comme si le *problème facile* de la conscience, celui qui consiste à comprendre le fonctionnement du cerveau en des termes scientifiques, était substitué par un habile escamotage au problème difficile de la conscience — qui consiste quant à lui à saisir pourquoi Je sens quelque chose, et comment se fait-il qu'il existe une correspondance quasi-magique entre les réponses de mon organisme et la sensation subjective de penser, de souffrir, de jouir,

bref, de vivre le monde de façon intimement personnelle.

Je n'ai encore parlé de ces chiasmes métaphysiques que comme s'il s'agissait de vérités avérées, sans en montrer la genèse ni la pertinence dans les théories contemporaines. C'est pourquoi deux interrogations s'imposent immédiatement : (1) le chiasme externaliste est-il présent dans les neurosciences ? et (2) quelle est son origine historique et épistémologique ? En particulier, on peut se demander s'il a toujours été admis que la conscience fût une propriété du cerveau. Dans le cas contraire, quand l'idée a-t-elle vue le jour ?

Afin de faire un peu de lumière sur ces questions, cette partie s'ouvrira sur une brève chronologie critique de l'idée d'après laquelle la conscience est logée dans le cerveau. Ce sera l'occasion d'exhiber quelques-uns des principes fondateurs des neurosciences contemporaines, tels que la méthodologie lésionnelle (étude des patients cérébrolésés), le principe de ségrégation fonctionnelle (chaque partie du cerveau possède un rôle) ou la plasticité neuronale (les parties du cerveau peuvent être réallouées). Cela soulèvera également le problème de savoir quels rapports les parties du cerveau entretiennent avec l'esprit lui-même ? Ou quels rapports entretiennent les organes des sens avec l'organisme dans son entier ?

Dans les chapitres ultérieurs, je me consacrerai plus spécifiquement à traquer la présence de chiasmes métaphysiques dans les théories de la conscience en vigueur aujourd'hui. Par « théories en vigueur », je n'entends pas nécessairement des théories ayant reçu l'adhésion scientifique unanime, ni même des théories qui sont actuellement utilisées. Je m'intéresserai plutôt aux théories ayant exercé, ou exerçant encore, une influence importante dans la façon de poser le problème de la conscience d'un point de vue neuroscientifique. Cette influence tient souvent plus à la nature des principes épistémologiques mobilisés qu'aux détails techniques et scientifiques des idées proprement dites.

Un chapitre intitulé « La Théorie darwinienne de Gerald Edelman » sera dédié à la théorie sélective des groupes neuronaux de Gerald Edelman. Ce prix Nobel de physiologie (1972) s'inscrit dans une critique du cognitivisme. Il introduit en outre une vision résolument émergentiste de la conscience, qui fait de celle-ci une propriété, non pas des neurones individuels, ni d'un groupe de neurone, mais de l'activité globale du cerveau. Je tenterai de montrer les avantages de sa stratégie non computationnelle (l'esprit n'est pas un ordinateur) et non dualiste (l'esprit n'est pas différent de la matière), tout en pointant du doigt l'existence d'un chiasme externaliste. Le danger de parler de pouvoir causaux et l'absence d'une analyse proprement phénoménologique de ce que c'est que « d'être conscient » sera passer au crible de ma critique.

Un autre chapitre, baptisé sommairement « La Théorie Intégrée de l'Information », s'intéressera à la théorie de Tononi et de ses collègues, qui prolonge les travaux d'Edelman dans une veine inspirée de la théorie de l'information de Shannon. Je suggérerai qu'il s'agit d'une des seules théories

neuroscientifiques à prendre au sérieux le versant proprement phénoménologique de la question de la conscience. Elle cherche en effet à tisser une analogie de structure entre les « axiomes » de la conscience, tels qu'ils se révèlent dans le « Je pense » introspectif, et les principes de fonctionnement fondamentaux du cerveau. Il sera néanmoins fait état d'une contradiction dans la description de la conscience vécue. Du reste, je critiquerai l'idée d'après laquelle la conscience est quelque chose d'immédiat et de représentationnel. J'attaquerai également l'opinion d'après laquelle la conscience soit un pur et simple traitement de l'information.

À l'issue de ce voyage dans le monde aussi prometteur que tendanciellement réducteur des neurosciences, je tâcherai d'isoler un certain nombre de principes épistémiques indispensables, d'après mes hypothèses de travail, à la construction d'une théorie moniste et bi-perspectiviste de la conscience, où le corps et la matière sont les deux facettes d'une même ontologie. L'on verra que les ingrédients clefs tiennent dans la méréologie (l'articulation du tout et des parties), l'analogie de structure (les lois de l'esprit épousent les lois de la matière), l'intégration (unicité des représentations dans l'acte de conscience) et la différenciation (la conscience se porte sur les changements).

13.1 La Conscience est-elle dans le cerveau ?

Le rôle régulateur de l'organe cérébral est un fait connu aussi vieux que l'histoire de l'humanité, ou tout au moins que l'entrée de l'humanité dans l'histoire : le papyrus d'Edwin Smith, datant de la VI^{ème} ou VII^{ème} dynastie, recense pas moins de 48 lésions, pour la plupart résultant de blessures de guerre, dont plus ou moins un quart concerne des atteintes de la boîte crânienne ou de la colonne vertébrale. On y trouve la première occurrence du nom « cerveau », mais aussi certains passages attestant d'une compréhension adéquate du système nerveux sympathique et parasympathique :

« Si tu procèdes à l'examen d'un homme ayant un désajustement d'une vertèbre du cou, et que tu constates qu'il n'a plus conscience de ses bras et de ses jambes à cause de cela, alors que son phallus est en érection à cause de cela, alors que de l'urine est tombée de son membre sans qu'il en ait eu conscience, alors que sa chair a reçu de l'air, alors que ses yeux sont injectés de sang : c'est un désajustement d'une vertèbre de son cou qui rejoint sa colonne vertébrale qui a causé qu'il est inconscient de ses bras et de ses jambes. »

C'est une chose, coriace s'il en est, de décrire pragmatiquement le fonctionnement du corps, mais c'en est une autre que de se représenter le lien

ténu qui semble exister entre la corporéité et l'origine de nos capacités sensorielles. En dépit d'un éclairage chirurgical assez précoce, de nombreux débats ont eu lieu concernant le rôle du cerveau dans les fonctions proprement sensibles, émotives et cognitives. Dans les balbutiements de la science médicale, qu'une certaine tradition aime à faire remonter à Hippocrate, deux visions s'opposent : l'encéphalocentrisme et le cardiocentrisme.

Si le père incontesté de la médecine, en visionnaire, reconnaît au cerveau plus qu'un simple régulateur de la température corporelle, ce n'est pas le cas de tous ses contemporains. Aristote et, dans une moindre mesure, Praxagoras, font du cœur le siège de la sensation. Il faut attendre Hérophile, Érasistrate et, plus tard, Galien de Pergame, pour que la vision encéphalocentrique gagne en importance et en précision, mais elle ne triomphera pas véritablement avant le XVI^{ème} et XVII^{ème} siècle, sous les travaux de Vésale ou de Thomas Willis. Même après que l'idée de la prévalence du cerveau se soit introduite dans le corpus médical, la théorie ventriculaire, selon laquelle ce sont les espaces vides du cerveau qui joueraient un rôle prédominant, garde les reins solides.

Encore aujourd'hui, il serait difficile de ne pas reconnaître l'existence d'une corrélation entre les fonctions mentales et le fonctionnement de l'organisme dans son entier : les pulsations du cœur, la réponse galvanique cutanée, la fréquence respiratoire, la température corporelle et le diamètre pupillaire ne sont-ils pas de bons prédicteurs de notre activité mentale et émotionnelle ? Si ce n'était pas le cas, pourquoi continue-t-on à les utiliser dans les détecteurs de mensonge ? S'il est vrai que les lésions du cerveau entraînent, selon toute apparence, une cessation de l'activité consciente — encore que nul ne puisse dire ce qu'il advienne de la conscience après la mort — on pourrait arguer, d'après les mêmes principes, que la cessation du cœur présente le même effet. Dès lors, pourquoi faut-il accorder un rôle prédominant à l'organe du cerveau dans la production présumée du phénomène de conscience ?

Les neurosciences ne semblent pas donner une réponse définitive à cette question, mais accumulent néanmoins un nombre important de preuves d'après lesquelles les parties du cerveau sont corrélées étroitement à certaines fonctions cognitives dont on estime, pour ainsi dire implicitement, qu'elles participent à la conscience. À cet égard, Thomas Willis peut être considéré comme le premier véritable précurseur d'une théorie du cerveau faisant la part belle aux zones corticales, sans oublier, entre autres mérites, l'importante distinction qui est faite entre la substance blanche et la substance grise. Dans son traité *Cerebri Anatomy*, il opère une taxinomie des nerfs crâniens et associe certaines fonctions cognitives au striatum ou au corps calleux. Les mouvements volontaires, par exemple, étaient localisés dans le cortex tandis que les mouvements involontaires étaient imputés au

cervelet^[1]. La neuroanatomie moderne, en hommage à ses travaux, a prêté son nom à la structure polygonale anastomotique présente à la base du cerveau.

À partir de ce point, la question centrale s'est incidemment transposée. Plutôt que de se demander qui du cœur ou du cerveau aurait l'apanage de la pensée, il s'agissait d'imputer les fonctions spécifiques de l'intelligence aux aires bien connues du cerveau, à la façon d'une carte corticale ou d'un vaste tableau de Mendeleïev de type neurophysiologique. Ce glissement est amorcé par les anatomistes Franz Joseph Gall (1758-1828) et Johann Caspar Spurzheim (1776-1832), les fidèles défenseurs de la phrénologie, aussi connue sous le nom de « théorie des bosses » en raison de sa prétention à associer des fonctions mentales à la forme du cerveau. En dépit d'un avenir prometteur en matière d'investigation criminelle, et du secours qu'elle reçoit de la part d'Auguste Comte^[2], cette théorie se heurte de plein fouet à la conception holistique du fonctionnement cérébral.

La principale objection scientifique à l'adresse de la phrénologie, si on écarte pour le moment les discussions de nature méthodologiques^[3], sont à trouver dans les travaux de Pierre Jean-Marie Flourens (1794-1897) sur les pigeons. Il est observé que ces animaux récupèrent progressivement des lésions cérébrales opérées expérimentalement. Ce qui semble prééminent, sur le plan comportemental, n'est pas tant la localisation de la résection, que la quantité de matière ôtée. Cela suggère de façon magistrale l'idée d'une certaine plasticité cérébrale, sans signer toutefois l'acte de décès de la phénoménologie, ou tout au moins de l'intuition localisationniste qui présidait à ses conclusions. Et pour cause : le concept de ségrégation fonctionnelle reste au cœur de la démarche neuroanatomique, comme va en attester deux découvertes postérieures.

Tout d'abord, Hughlings Jackson (1835-1911), en se penchant sur des patients épileptiques, découvre que la manifestation des convulsions suit une progression bien spécifique dans le corps humains, comme si à chaque excitation cérébrale localisée correspondait, de façon quasi bijective, une partie du corps humain. Il va ainsi postuler l'existence d'une somatotopie fonctionnelle, anticipant en cela l'homoncule de Wilder Penfield du gyrus précentral et l'homoncule sensitif du cortex somesthésique^[4].

Dans la même lignée, le cas clinique désormais célèbre de Monsieur Leborgne, connu aussi sous le nom de Monsieur « Tan-Tan » en référence à

1. ARRÁEZ-AYBAR, LUIS-A et al., *Thomas Willis, a pioneer in translational research in anatomy (on the 350th anniversary of Cerebri anatome)*, Journal of Anatomy, 226 (3) : 289-300.

2. Cf. COMPTE AUGUSTE, *Système de politique positive*, Hermann, 2022.

3. Voir par exemple le chapitre de Hegel sur la phénoménologie.

4. Cf. PENFIELD W., BOLDREY E., *Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation*, Brain, 1937 et PENFIELD W., RASMUSSEN T., *The cerebral cortex of man*, New York : Macmillan, 1950

l'unique onomatopée qu'il était capable de produire, va permettre à Paul Broca (1824-1880) de localiser l'air de production du langage quelque part au pied de la troisième circonvolution frontale gauche, dans une région aujourd'hui connue sous son nom. Non content de poser ainsi les jalons de la neuropsychologie, cet événement sera suivi de très près par la découverte de Carl Wernicke (1848-1804), selon laquelle il existerait aussi une aire du cerveau dédiée, non à la production, mais à la compréhension du langage. Partant d'un ensemble d'observations cliniques opérés sur des patients cérébrolésés, Wernicke parvient à localiser cette fonction dans le lobe temporal gauche, mais aussi à prédire l'existence d'autres sortes d'aphasie, comme l'aphasie globale ou de conduction, ébauchant à sa manière le rêve de faire du cerveau une sorte de table de Mendéléev^[5].

13.2 Ségrégation fonctionnelle et plasticité neuronale

Aujourd'hui, bien que le holisme ait été globalement abandonné, il semblerait que ces deux notions, celle de ségrégation fonctionnelle et de plasticité neuronale, soient conciliées de façon cohérentes dans les représentations du cerveau. Je parle bien de représentations au plurielle car la neuro-anatomie des commencements, suivant en cela son accroissement naturel, a donné naissance à toute une pléthore de sciences nouvelles : citons entre autres la neuro-histologie, la neuroendocrinologie, la neuro-physiologie, la neurologie, la neuropsychologie, l'électrophysiologie, ou encore la neuro-pharmacologie. Aussi diverses que soient ces disciplines neurobiologiques, ils semblerait qu'elles puissent se ranger dans quatre grandes approches : l'approche top-down, l'approche bottom-up, l'approche corrélation structure-fonction et, enfin, celle des systèmes organisés^[6].

L'hypothèse commune à toutes ces approches, c'est qu'il est possible d'expliquer la vie mentale intime d'un individu, considéré du point de vue du « Je », à partir d'une théorie adéquate du cerveau. Tout se passe comme si le vieux débat entre encéphalocentrisme et cardiocentrisme n'avait fait que changer de nature, sans abandonner à l'histoire ancienne ses préoccupations originaires. Si tous s'accordent à dire que des expressions linguistiques aussi variées que « apprendre par cœur », « avoir le cœur brisé » ou « avoir le cœur sur la main » ne sauraient être que métaphoriques, on est en droit de se demander ce qu'il en est des expressions comme « être cerveau du complot », « fuite des cerveaux » ou encore « faire tourner ses méninges » ? Sans tomber dans le travers des vulgarisateurs de comptoir, qui irai●t dire

5. WERNICKE C, *Grundriss der Psychiatrie in klinischen Vorlesungen*, 1900.

6. Cette catégorisation est empruntée à MONIER CYRIL, *Les neurosciences au sein des sciences de la cognition : Une revue de la littérature plaidant pour un devenir spinoziste des neurosciences*, 2019.

que « tomber amoureux, c'est chimique^[7] », il y a une certaine tendance à faire du cerveau quelque chose d'analogue à ce que les anciens faisaient du cœur humain : à savoir, le centre de la pensée et des émotions.

Il est important de bien distinguer, à ce stade, la véritable pratique neuroscientifique de la caricature que ses adversaires en ont fait. Il serait trop facile, du point de vue qui est le nôtre, de s'en prendre à un épouvantail com-mode^[8]. Une certain pan de la littérature dite « neurosceptique » a beau jeu de nous dépeindre le neuroscientifique comme une sorte de philosophe naïf, un réductionniste de pure souche qui affectionne les phrases grandiloquentes sur le statut, oh combien supérieur, de la science^[9]. Or lorsqu'on regarde la pratique, il semblerait que l'approche « bottom-up », cherchant à expliquer les fonctions cognitives d'ordre supérieur en partant des neurones, ne soit pas la stratégie la plus dominante en neurosciences. D'après Daniel Stoljar et Ian Gold, une grande majorité des neuroscientifiques se situeraient dans le paradigme des systèmes émergents^[10]. Ils défendraient dès lors une forme de réductionnisme faible, quand ils ne feraient pas tout simplement un aveu d'ignorance métaphysique.

Peut-être est-il judicieux d'adhérer à l'idée selon laquelle la conception réductionniste en neuroscience, que certains auteurs nomment la doctrine du neurone^[11], peut être entendue de deux manières différentes ? En suivant l'article de Daniel Stoljar et Ian Gold, l'affirmation que seule les neurosciences cognitives puissent rendre compte du fonctionnement de l'esprit pourrait dès lors se comprendre de deux façon bien différentes :

- De façon triviale : en prenant le terme « neurosciences cognitives » dans son acception la plus large. Il s'agit alors de cette approche interdisciplinaire que l'on situe traditionnellement au croisement de la biologie et de la psychologie, qui prescrit certes une méthodologie et une finalité, mais non une théorie spécifique. Les seules présuppositions, selon Stoljar et Gold, sous-jacentes à cette version triviale de la doctrine, sont de trois types. Tout d'abord, on observe une adhésion au matérialisme, dans la mesure où l'esprit ne relève pas d'une autre substance que celle du monde physique, bien qu'il puisse y avoir ici présence d'émergence ou de supervénience. Ensuite, on

7. VINCENT L., *Comment devient-on amoureux ?*, Paris, Odile Jacob, 2004.

8. Cf. RAMUS F., *L'épouvantail com-mode*, 2014.

9. Un exemple de telles assertions malheureuses peut être trouvé dans l'ouvrage de Crick : « Vous, vos joies et vos peines, vos souvenirs et vos ambitions, votre perception de l'identité personnelle et du libre arbitre, ne sont, en fait, rien de plus que le fonctionnement d'un vaste assemblage de cellules nerveuses et de leurs molécules associées. Vous n'êtes rien d'autre qu'un paquet de neurones. » (CRICK, F., *The Astonishing Hypothesis*, New York, Charles Scribner's Sons, 1994).

10. GOLD, I, STOLJAR D, *A neuron doctrine in the philosophy of neuroscience*, Behavioral and Brain Sciences, 22(5), 809-30 ; discussion 831-69.

11. *Ibidem*.

trouve l'idée que l'esprit puisse être décrit de façon correcte par une théorie scientifique, ce qui revient à une forme de naturalisme. Enfin, il est dit que la biologie seule est insuffisante en l'état, qu'elle nécessite le recours possible à des concepts issus de la psychologie. À titre d'exemple, on pourrait envisager que la psychologie prescrive des explications d'ordre fonctionnelles tandis que la biologie traiterait des mécanismes.

- De façon radicale : en substituant au terme de « neurosciences cognitives » celui de « neurosciences biologiques ». Dans ce cas, la doctrine devient beaucoup plus ambitieuse, puisqu'elle implique que les concepts utilisés en linguistiques ou en psychologie cognitive ne sont que des expressions simplifiées et redondante du phénomène qui est en cause, et qui est, lui, de nature neurobiologique. Une théorie de la mémoire de travail qui parlerait de « boucles phonologiques », de « calepin visuo-spatial » ou de tampon épisodique » ne serait donc qu'une explication parabolique, ou disons un point de relais, avant de s'élever vers la véritable compréhension du phénomène¹². Dans cette vision radicale, il est dit quelque chose de plus que l'idée que l'esprit soit scientifiquement descriptible, mais cela se fait au coût d'une perte de vraisemblance.

Cette distinction a donné naissance à de nombreuses polémiques, que nous analyserons en temps utile. Ce qui nous souhaitons relever en guise de préliminaire, c'est que même dans la conception triviale de la doctrine du neurone, qui de l'aveu des auteurs semblent partagée par tous les neuroscientifiques, le cerveau reste au centre des préoccupations. Il n'est pas seulement central au sens où il serait l'organe le plus important de l'être humain, c'est-à-dire ce sans quoi l'être humain ne peut plus être pensée ; il est central en ce sens que des relations explicatives semblent pouvoir être dressées entre les mécanismes biologiques du cerveau et des vécus psychiques relevant de l'expérience de conscience directe. Ce que nous avons en vue ici, c'est la prétention des neurosciences à combler l'écart existant entre le point de vue à la troisième personne et le point de vue à la première personne. Si en effet nous voulons soutenir, comme annoncé dans l'introduction, un monisme à double aspect, il est important de répondre à cette importante objection adressée par la science, qui ferait de l'organe humain une passerelle entre les deux points de vue pronominaux.

Dans leurs articles original, Stoljar et Gold n'envisagent que trois sortes d'alternatives au naturalisme matérialiste induit par la doctrine triviale du

12. Je pense ici notamment à l'article de Kim (1992, 1999), qui a bien systématisé ce problème, et qui a donné naissance à toute une lignée de théories émergentistes. Pour une taxonomie plus étendue, voir aussi Sartenaer Olivier (2015).

neurone. La première consiste à rejeter l'hypothèse que (a) l'esprit est un phénomène biologique. Ils donnent pour exemple le dualisme, qui semble aujourd'hui trouver encore pas mal de partisans. La seconde approche consiste à nier l'hypothèse que (b) l'esprit soit l'objet d'étude des sciences naturelles. Ils illustrent cette possibilité en invoquant les théories du constructivisme social, mais en rappelant que le dualisme est aussi un très bon candidat. La troisième stratégie, quant à elle, consiste à récuser l'idée comme quoi (c) la biologie et la psychologie sont, en principe, suffisants pour expliquer l'esprit humain. Un exemple de théorie qui se range à ce niveau seraient le programme de l'intelligence artificielle forte, dans sa forme la plus puritaine, en ce qu'elle envisagerait une théorie de l'esprit en des « termes abstraits^[13] », qui n'aurait pas besoin du secours de ces deux disciplines.

13.3 Spinozisme neuroscientifique

Nous souhaitons soutenir l'existence d'une quatrième voie. Nous la pensons viable, sur le plan logique, et compatible avec la pratique des neuroscientifiques eux-mêmes. Cette approche consisterait à réfuter que (a') l'esprit serait uniquement compréhensible de façon satisfaisante lorsqu'on l'envisage comme un phénomène biologique et que (b') l'esprit serait uniquement l'objet d'étude des sciences naturelles. La présence du terme « uniquement », que nous avons mis en exergue, n'a pas pour vocation de renvoyer à un dualisme dissimulé. Il réfère plutôt à la métaphore qui a été ébauchée en introduction : de la même façon qu'un objet quantique peut être envisagé à la fois comme une particule et à la fois comme une onde (principe de complémentarité de Bohr), nous souhaitons défendre l'idée que l'esprit puisse à la fois se comprendre sous le mode phénoménal et sous celui de la matière^[14].

Loin d'imaginer que l'esprit soit une substance volatile piégée dans la matière, à la façon d'un fantôme dans une machine, ou que l'esprit ne soit rien d'autre que de la matière en interaction, nous suggérons que l'esprit et la matière sont deux manières complémentaires, quoique mutuellement exclusives, de nous représenter le monde qui nous entoure — et nous-même qui l'observons par la même occasion. Si la perspective « idéaliste » tire son origine dans le point de vue à la première personne (la perspective en « Je »), la perspective « matérialiste », quant à elle, trouve sa source dans le point de vue à la troisième personne (la perspective en « Il »). La première a pour

13. Ce sont les termes utilisés par les auteurs : “An example of the third kind of theory would be a certain version of the artificial intelligence program, according to which both neurobiology and the details of psychology are in principle irrelevant to the construction of theories of mentality in the most abstract sense” (Stoljar et Gold, *op. cit.*, page 814).

14. Notons qu'il s'agit d'une métaphore concernant l'accès à la connaissance, et non pas la nature du fonctionnement neuronal ou de la conscience en tant que telle. Il n'est pas stipulé quoi que ce soit à propos de la pertinence de la physique quantique pour résoudre certains problèmes se présentant en philosophie de l'esprit.

Pierre de touche ce que Husserl appelait la régression phénoménologique transcendantale. La seconde se fonde sur la stabilité du monde qui nous entoure et la capacité que nous avons à le prédire de façon pragmatiquement efficace. En l'état, il est difficile de prouver, à l'aide d'un argument définitif, l'existence d'une réalité extérieure à notre vécu phénoménal, comme il est difficile d'expliquer aisément le miracle de la science autrement qu'en recourant à l'idée d'un monde tangible. Ces deux représentations semblent donc viables.

Cela ne veut pas dire pour autant qu'il existerait comme deux substances distinctes. Nous disons que le monde est à la fois envisageable sous deux modalités, et non pas en même temps. Le problème central du dualisme est précisément la difficulté de rendre compte de l'articulation, des relations causales, existant entre la matière et l'esprit¹⁵. Ce problème n'existe pas sous notre hypothèse, puisque le dualisme n'est pas à prendre dans un sens ontologique. Les deux modalités ne sont que des points de vue, humains, posés sur une substance, unique, dont les informations coïncideraient au moins partiellement ; de sorte que, du point de vue en « Il », presque tout peut s'expliquer de façon matérialiste, tandis que du point de vue du « Je », presque tout peut s'expliquer au sein-même de la conscience. Les difficultés n'apparaissent qu'aux points de rencontre entre ces deux points de vue, aux frontières respectives des deux sphères explicatives. Il en va ainsi de la question de savoir comment un agrégat de matière peut avoir une vie intime, propriété désignée tantôt sous le nom de qualia, tantôt sous le nom de « Je » phénoménal. Il en est aussi de même lorsque l'on se demande, au sein d'une approche à la première personne, pourquoi il existe une liaison si forte entre certaines fonctions cognitives, comme l'exercice du langage, et des aires déterminées du cerveau (l'aire de Broca ou de Wernicke).

Dans un cas comme dans l'autre, le problème vient de ce que l'on assemble, dans une même question, le vocabulaire spécifique à ces deux points de vue, qu'on cherche en quelque sorte à en faire le chiasme. La question de savoir si cette contradiction vient de ce que notre point de vue humain est incapable d'appréhender le réel dans toute sa complétude, ou si elle correspond au contraire à quelque chose de plus fondamental dans la nature, est plus spéculative. Elle sera traitée dans la deuxième partie de cette thèse, conjointement avec toutes sortes d'autres objections. À l'heure actuelle, nous ne faisons qu'esquisser l'hypothèse méthodologique selon laquelle il y a quelque chose qui est dit, dans le point de vue à la première personne, qui, méthodologiquement, ne peut se réduire au point de vue à la troisième personne ; et inversement.

Il doit être remarqué, avant d'aller plus loin, que ces deux points de

15. Je pense ici notamment à l'article de Kim (1992, 1999), qui a bien systématisé ce problème, et qui a donné naissance à toute une lignée de théories émergentistes. Pour une taxonomie plus étendue, voir aussi Sartenaer Olivier (2015).

vue, celui en « Je » et celui en « Il », présentent une propriété commune : l'auto-consistance. Bien qu'ils soient incomplets, l'un comme l'autre se suffisent à eux-même et n'ont pas besoin, en principe, de l'autre pour être conceptualisé. Il y a certes des phénomènes qu'ils ne peuvent pas expliquer de façon satisfaisante, et qui sont comme les pierres d'achoppement, mais ceux-ci n'apparaissent que lorsqu'on tente d'expliquer l'un par l'autre. Et pour cause : aucune tentative ne peut être faite d'expliquer l'esprit à partir de la matière (réification) ou la matière à partir de l'esprit (idéalisat-ion). Des énoncés du type « l'activité de ce groupe neuronal participe de la conscience » présenteraient une contradiction dans les termes, à moins de ne prendre le mot « conscience » comme autre chose qu'une manifestation matérielle objective. Que voulons-nous dire par là ? Il est certes possible de faire une expérience consistant à mettre un sujet en présence d'un encéphalogramme et d'étudier les relations existants entre les signaux mesurés et les perceptions conscientes auto-rapportés. Il est aussi possible d'étudier ce que deviennent les perceptions auto-rapportées par suite de lésions cérébrales spécifiques. Ce genre d'expérience permet d'établir une corrélation certaine entre l'activité neuronale et le vécu psychologique. Elle renforce l'idée que le cerveau est une condition nécessaire de la pensée, non en ce qu'il la cause, mais en ce que ces deux événements que sont (a) les activités neuronales et (b) la perception consciente auto-rapportée (un fait comportemental), surviennent de façon concomitante. Franchir le pas qui consiste à mettre la matière en relation avec, non plus les manifestations comportementales de conscience, mais la conscience elle-même, en ce qu'elle a de vécu, est aussi dangereux que de franchir le fossé de l'induction humienne qui sépare la vérité apodictique de la simple vraisemblance. L'existence d'une corrélation nous pousse certes à croire en l'existence d'une réalité sous-jacente, mais nous fait sortir du même coup de la scientificité.

Lorsqu'en effet nous nous représentons l'esprit comme quelque chose d'étendu, nous perdons de vue sa propriété essentielle. Nous faisons de lui un objet tangible indépendant de l'observateur, et non pas un vécu. À contrario, lorsque nous parlons de la matière comme d'un acte de conscience, à la façon de l'idéalisme subjectif, nous le réduisons tout simplement à sa manifestation phénoménale. Il n'y a pas plus de sens à chercher à expliquer comment la pensée peut provenir des neurones, que de se demander s'il existerait une substance pensante logée quelque part dans nos cellules. Cela veut-il dire que chercher le lien existant entre les deux est une voie sans issue ? Tout dépend la façon de procéder. Si en effet la pensée et la matière sont les deux facettes d'une même réalité, il est vraisemblable de croire que cette réalité se comporte d'après des lois identiques. Il serait dès lors possible d'exhiber l'existence d'une structure commune entre les lois de l'esprit et les lois de la matière. Cette structure nous aiderait à comprendre comment s'articulent, entre eux, les inputs neuronaux (perspective en « il ») mais aussi comment s'articuleraient entre eux les vécus phénoménaux (perspective en

« Je »). Une telle structure ne représenterait aucunement une preuve, mais elle est nécessaire à la viabilité de l'hypothèse elle-même. C'est la raison pour laquelle il est important d'isoler les principes épistémiques présidant à la formation de l'esprit dans le paradigme neuroscientifique et de voir s'il est compatible avec ceux que l'on découvrirait dans une investigation à la première personne, afin de falsifier, le cas échéant, notre hypothèse.

Bien que cette hypothèse d'une double modalité puisse paraître étrangère aux neurosciences, d'un point de vue épistémique, elle est pourtant défendue par plusieurs grands noms dans le domaine. Et pour cause, son inspiration est étroitement liée au spinozisme :

« L'esprit n'émerge pas de la matière comme s'il était causé par elle, de même qu'il ne cause pas le corps dont il est l'idée, ni aucun autre. Il y a émergence de la complexité matérielle du cerveau à partir des éléments plus simples – les neurones et leurs connexions – qui le constituent. Et il y a émergence de l'esprit, idée du cerveau (ou, plus précisément, du corps) à partir des idées des neurones, étant entendu [...] que l'idée d'un neurone, comme d'un électron ou d'une pierre, n'est pas autre chose que sa loi d'organisation et de fonctionnement. Celle-ci est plus simple, bien que relevant aussi de l'attribut Pensée que l'idée du corps humain, c'est-à-dire l'esprit humain. Les propriétés de celui-ci émergent de l'ensemble des interactions entre les idées des constituants du corps. L'esprit et le corps humain ont ainsi tous les deux des propriétés émergentes de leurs constituants. Mais ils sont tous deux, ainsi que leurs constituants respectifs, deux aspects de même chose vues sous un attribut ou un autre. » (Atlan, 2018).

La stratégie des prochains chapitres est donc double. D'une part, il s'agira de montrer qu'il existe bel et bien, dans les travaux des neuroscientifiques, une confusion entre le point de vue à la première personne et le point de vue à la troisième personne. En dépit de certaines précautions, il est souvent dit quelque chose de plus, à propos du cerveau, que ce qu'on en dit généralement à propos des autres organes. Loin d'être vu comme une simple condition nécessaire de la pensée, ou comme des phénomènes qui coïncident avec les propriétés conscientes auto-rapportés des sujets, les modèles neuro-naux ont généralement une prétention à capturer, au moins partiellement, le versant subjectif de l'expérience consciente.

Dans un deuxième temps, il s'agira d'isoler les principes épistémologiques qui rendent possibles la connaissance de l'esprit et montrer qu'ils sont compatibles avec le monisme à double aspect, ou tout au moins le spinozisme neuroscientifique, que j'entends défendre dans cette thèse. Les conclusions qui en seront tirées serviront de fil conducteur pour l'élaboration ultérieure d'une théorie méréologique de la conscience.

Il existe dans la pratique scientifique une confusion entre le point de vue en « Je » (conscience subjective) et le point de vue en « Il » (conscience objective). Cette confusion est ce que j'appelle :

- un *chiasme métaphysique externaliste*, lorsque le « Je » est réduit au « Il ».
- un *chiasme métaphysique internaliste*, lorsque le « Il » est réduit au « Je ».

Ces chiasmes peuvent être évités si, au lieu de vouloir réduire la conscience vécue à une théorie biologique du cerveau, on admet que la matière et la conscience sont deux perspectives posées sur une même réalité.

J'ai montré que ce monisme à double aspect est proche du spinozisme neuroscientifique et trouve de nombreux défenseurs dans la communauté scientifique. Je suggère en outre que, si la conscience et la matière sont les deux visages d'une même réalité, ils doivent partager une même structure.

Cela veut dire qu'il doit être possible de tisser une analogie entre les principes régissant le fonctionnement de la conscience (envisagées comme un « Je » introspectif) et les principes régissant le fonctionnement du cerveau (envisagé comme un « Il » observationnel).

Les chapitres suivants vont étudier deux théories neuroscientifiques du cerveau, celle d'Edelman et celle de Tononi.

Chapitre 14

La Théorie darwinienne de Gerald Edelman

Rien ne fait sens en biologie sans la lumière de l'évolution.

Theodosius Dobzhansky

Il n'est pas rare qu'un éminent scientifique, sitôt qu'il s'est vu octroyé le prix Nobel — et poussé en ce sens par le monde médiatique — se pique soudain de parler tout azimuts. Qu'il s'agisse de politique, d'économie, du bonheur, de la vie extraterrestre ou de la conscience, tout se passe comme si l'exercice des sciences naturelles, pour des raisons qui n'auraient rien à envier à la théorie de la grâce révélée de Saint-Thomas d'Aquin, disposait l'expert à s'emparer, sans aucun préavis, du ciel entier des idées platoniciennes.

De tels phénomènes suscitent généralement une sainte méfiance, mais lorsque le glissement se cantonne à un sujet ciblé, allant par exemple de l'immunologie à la conscience, ou de la biologie à l'épistémologie, et lorsqu'il se produit dans le chef d'un savant tel que Gerald Edelman, qui allie un certain talent vulgarisateur à la circonspection philosophique, cela peut aboutir à des axes de recherche intéressants^[1]. En cette ère d'accélération sans précédent, la science n'a peut-être rien besoin le plus au monde que d'une approche globale, synthétique, multifactorielle, capable de transposer, par un habile jeu de vases communicants, les connaissances d'un domaine vers un autre, de sorte à réunir, à rapprocher, à éventer — et dépoussiérer au passage — les disciplines traditionnelles.

1. Dans le discours adressé à l'adresse du comité Nobel lors de l'obtention de son prix en 1972, Gerald Edelman n'hésite pas à formuler explicitement l'idée d'après laquelle : « Immunology is particularly exciting, however, because it provokes unusual ideas, some of which are not easily come upon through other fields of study. Indeed, many immunologists believe that for this reason, immunology will have a great impact on other branches of biology and medicine. » (EDELMAN GERALD M., *Antibody Structure and Molecular Immunology*, Nobel Lecture, December 12, 1972)

Gerald Edelman s'est fait connaître pour ses recherches inaugurales sur la structure des immunoglobulines, ou lymphocytes B, pour lesquelles il a reçu le prix Nobel de médecine en 1972 avec son collègue Rodney Porter. L'objet de ce succès ? Il est parvenu à déterminer les séquences d'acides aminés des chaînes d'anticorps monoclonaux de patients souffrant d'un syndrome immunoprolifératif. Non content de cela, il a cherché à répondre à la question de savoir comment une structure générale des lymphocytes peut donner naissance à autant de variantes d'anticorps, adaptées à chaque molécule d'antigène ? À cette même époque, il a développé le concept audacieux selon lequel plusieurs gènes contrôlent la synthèse des molécules d'anticorps, dont certains sont responsables de la partie constante de la molécule et d'autres de la partie variable. Ces travaux pionniers seront ultérieurement confirmés par la biologie moléculaire.

Après avoir remporté le prix Nobel en 1972, G. Edelman a tenté de transposer les résultats de ses théories immunologiques à l'étude du cerveau et de la conscience : « Ces mécanismes fondent ce que j'ai appelé la théorie de sélection des groupes neuronaux [...] Tous les processus à l'œuvre dans le développement et la morphogenèse du cerveau présentent un certain côté statistique : la neurogenèse, la migration neuronale, l'adhésion qui régule la croissance des extrémités neuronales, la mort neuronale programmée, la formation des connexions [...] Ces variations se combinent pour donner une diversité de structures absolument extraordinaire qui fait que deux individus, même jumeaux, n'ont jamais le même cerveau. C'est la sélection développementale, premier pilier de ma théorie. Par la suite le câblage entre groupes neuronaux va être modifié en fonction des interactions avec l'environnement, notamment sensoriel²... »

Dans ce chapitre, j'aimerais donner un aperçu de la théorie de la conscience élaborée par Edelman, aussi bien quant à ses aspects scientifiques qu'épistémologiques. Je tenterai de montrer ensuite la présence d'un chiasme métaphysique externaliste, consistant à mélanger des termes observationnels, se référant à des phénomènes objectifs de l'espace, à des termes introspectifs qui renvoient à l'expérience d'un sujet éprouvant le monde. Je conclurai en donnant quelques pistes d'élaboration d'une théorie biperspectiviste de la conscience.

14.1 Critique edelmanienne du dualisme

Avant d'aborder la théorie de la conscience d'Edelman, j'aimerais tenter d'identifier les intuitions épistémologiques qui ont guidé sa démarche. Edelman était-il un partisan du réductionnisme matérialiste ou du dualisme ? Pensait-il qu'il existait deux substances distinctes (le corps et l'esprit) ou

2. Cité depuis l'Académie des Sciences de France, [https ://www.academie-sciences.fr/fr/In-memoriam/gerald-maurice-edelman.html](https://www.academie-sciences.fr/fr/In-memoriam/gerald-maurice-edelman.html)

croyait-il que la conscience puisse, d'une manière ou d'une autre, se réduire partiellement à un ensemble de lois physico-chimiques ? Je vais commencer par discuter les objections qu'Edelman adresse à l'idée d'une substance distincte de la matière, après quoi j'étudierai, à rebours de cette opinion, les objections qu'il adresse à l'idée que la conscience fonctionne comme un ordinateur. Cela ouvrira la voie à une troisième approche, dite émergentiste, où la conscience, bien que supportée par des processus matériels, ne se réduit pas purement et simplement à des opérations neuronales.

Dans son ouvrage intitulé *Seconde nature : la science du cerveau et le savoir humain*, Gerald Edelman donne d'emblée la couleur de ses convictions métaphysiques. Il reproche au dualisme cartésien, en ce qu'il opère une distinction entre la *res extensa* (la matière) et la *res cogitans* (la pensée), d'avoir retardé l'étude scientifique de l'esprit humain.

*Un facteur ayant affecté l'approche scientifique de la conscience peut être imputé à la pensée très influente de René Descartes. Quelques temps après Galilée, Descartes a retranché l'esprit de la nature. Il fit cela par la pensée seule, concluant qu'il existait deux substances distinctes : la res cogitan, les choses étendues tombant sous la physique, et la res cogitan, les choses pensantes n'ayant pas d'étendue dans l'espace et qui n'étaient pas accessibles à la physique*³

La *res cogitan* défendue par Descartes serait, d'après Edelman, une substance réfractaire à la lunette du scientifique. Dans la mesure où cet ordre de réalité n'est accessible que via la perspective du « Je pense », elle ne peut pas faire l'objet d'une objectivité établie. C'est ce qui lui fait dire que l'esprit humain a été retranché de la nature. On comprend son désarroi. Dans la vision de Descartes, la conscience n'est pas plus accessible à l'observation scientifique que les planètes lointaines ne sont visibles à l'œil nu. La réalité du mental est reléguée du côté du monde invisible des phénomènes intimes.

S'il est compréhensible qu'Edelman s'attaque à la *res cogitan*, qu'il voit comme un résidu du concept d'âme, il est néanmoins légitime de se demander dans quelle mesure la conscience a résisté aux théories scientifiques, non pas en raison de préjugés, mais parce que cette conscience ne se soumet pas à la méthode expérimentale ? La science en effet, de ce qu'elle prend un point de vue en « Il », n'est pas à même de rendre compte de la dimension proprement phénoménologique, ou subjective, du monde mental. Le mérite de Descartes, loin d'avoir fait reculer la science, est d'avoir cherché à réunir la dimension proprement subjective et objective de la pensée.

La façon dont il procède, sur le pan épistémologique, est certes discutable. Il n'en reste pas moins que Descartes s'appuie sur des notions scientifiques établies à son époque : je veux parler, d'une part, des esprit animaux

3. EDELMAN G., *Second nature*, Yale University Press

et, d'autre part, de l'épiphyse. Cette dernière, plus connue sous le nom de « glande pinéale », se situe à l'entrée du canal de Sylvius. Elle est aujourd'hui reconnue comme responsable de la création de mélatonine, participant de la régulation du sommeil. Bien qu'elle ne joue pas aujourd'hui de rôle particulier dans les théories neuroscientifiques de la conscience, Descartes y voit un argument de symétrie : il s'agit du seul organe cérébral connu à n'avoir pas de duplicata dans l'hémisphère droit et dans l'hémisphère gauche. Elle occupe, comme le nez au milieu du visage — ou comme le trône au milieu du crâne — une position centrale. Au contraire des yeux et des oreilles, qui apportent sur un même objet deux images ou deux sons distincts, la glande rassemble selon Descartes ces diverses impressions sous une perception unitaire. C'est ce qui explique sa fonction intégrative, propre à la conscience⁴.

Les idées de Descartes, sur certains points d'anatomie, constituent une régression des connaissances établies depuis les travaux de Galien. Lorsqu'il affirme, par exemple, que la glande pinéale est suspendue au milieu du ventricule, et non pas à l'extérieur ; ou que cette glande est composée d'une substance volatile, et non liquide, il commet des erreurs scientifiques. Il semblerait donc en effet que le recours à la notion de l'âme ait participé à la popularisation d'une vision quelque peu magique de l'anatomie cérébrale. En cela, il faut donner raison à Gerald Edelman. Il est toutefois important de distinguer, pour rendre à Descartes les lauriers qui lui sont dus, le contenu des idées de l'intention sous-jacente de la démarche.

La persistance d'Edelman à s'attaquer au dualisme, en dernière analyse, ne s'appuie peut-être pas tant sur les erreurs scientifiques de Descartes que sur une possible confusion entre ce que Chalmers appelait le problème facile et le problème difficile de la conscience ? Il est relativement « facile », tout au moins en principe, de construire une théorie permettant d'expliquer comment le cerveau intègre l'information donnée par les sens pour produire des réactions évolutivement adaptées aux contraintes de l'environnement. Il est également « facile » de mettre en relation les mécanismes du cerveau avec les phénomènes autorapportés de la douleur, de la pensée ou de l'imagination. Il est en revanche difficile d'expliquer en quoi le fait d'être un « Je » pensant le monde est produit par le cerveau :

*Il est de première importance de prendre la conscience au sérieux.
La manière la plus simple de développer une « théorie » de la
conscience est de nier son existence, ou de redéfinir le phénomène
nécessitant une explication comme quelque chose qu'il n'est pas.
Cela conduit généralement à une théorie certes élégante, mais*

4. Cette théorie de Descartes sur la glande pinéale n'est pas une innovation en soi. On trouve chez Galien des affirmations d'après lesquelles les ventricules connexes à cette glande étaient remplies de *pneuma*, une substance volatile considérée par les anciens grecs comme synonyme de l'âme. Dans l'antiquité, on attribuait d'ailleurs à différents appendices du *cerebellum* un rôle de régulation du *pneuma*.

qui est loin de faire disparaître le problème⁵

14.2 Critique edelmanienne du computationnalisme

En dépit de sa critique du dualisme, Edelman ne défend pas un réductionnisme physicaliste aveugle. Il ne se représente pas le cerveau comme un ordinateur chargé de traiter l'information incidente en vue de produire, comme par effet de bord, des états mentaux. Il insiste au contraire sur le fait que l'esprit est incarné et le corps lui-même est intégré dans un environnement plus large. Chaque fois qu'il parle d'une propriété du cerveau, il précise qu'il s'agit d'une simplification pour désigner l'ensemble des signaux échangés entre les organes, dont le cerveau occupe une place centrale en tant que régulateur. Les échanges entre le corps et le cerveau, à leur tour, sont conditionnés par un certain rapport à l'environnement, formant ce qu'on appelle une éconiche⁶. À ce titre, Edelman inscrit sa pensée dans le paradigme énaïvistique dont nous avons parlé dans l'introduction.

Au-delà de l'accent porté sur le corps et l'environnement, Edelman critique explicitement la vision réductionniste (ou *reductio ad computatum*)⁷ d'après laquelle l'ordinateur devrait servir de modèle au fonctionnement du cerveau, et ce pour trois raisons fondamentales. Premièrement, rappelle Edelman, un ordinateur recourt à l'arithmétique et au langage symbolique. Il manipule des objets dans de courts intervalles de temps, régulés par une horloge interne. C'est-à-dire qu'il existe des unités de temps prédéfinies, au cours de chacune desquelles peut être lue une instruction du type : « soustraire la grandeur a à la variable b », « assigner la valeur a à la variable b », etc. Edelman défend toutefois que le cerveau n'est pas régulé par des règles logiques, mais, comme je vais le montrer par la suite, selon un fonctionnement émergent de l'association sélective des groupes neuronaux.

Du reste, un ordinateur, pour fonctionner correctement, doit recevoir un input non ambigu. Il ne travaille pas avec des grandeurs analogiques, mais bien des entiers naturels. Or les signaux dont s'occupe le cerveau ne sont pas bien organisés : loin de s'apparenter à des caractères sur un ruban, le monde physique, le monde tangible de la vie de tous les jours, ne présente pas de structure intrinsèque. Tout le problème de l'intelligence artificielle consiste d'ailleurs à organiser les données sensibles en catégories commodément à manipuler. C'est d'ailleurs pourquoi que la première instance célèbre de

5. La Traduction est de moi : « *The first and most important is to take consciousness seriously. The easiest way to develop a "theory" of consciousness is to deny its existence, or to redefine the phenomenon in need of explanation as something it is not. This usually leads to an elegant theory, but the problem does not go away.* », DAVID J. CHALMERS, *The Conscious Mind*.

6. GERALD M. EDELMAN *Second Nature : Brain Science and Human Knowledge*, p.25

7. Cette expression est de moi. Elle fait le lien avec ce que j'ai dit dans l'introduction de la première partie.

réseaux de neurones, qui sert aujourd'hui de pierre de touche à tous les pédagogues, consistait à lire des chiffres écrit par voie manuscrite. Il s'agissait donc de prendre en input un ensemble de caractères visuels aussi vastes que complexes, composés de points, de lignes, de ratures, de traits aux formes plus ou moins épaisses, de les réduire à un ensemble de pixels identifiables, puis d'associer ces groupes de pixels aux chiffres 0, 1, 2,... En des termes kantien, on pourrait dire que la spécificité du cerveau est subsumer le contenu sensible, cette rhapsodie d'images, de sons, de couleurs, sous un concept prêt à l'emploi.

La deuxième critique d'Edelman consiste à souligner que la structuration des réseaux neuronaux suivent un processus propre à chaque individu. « Il n'existe pas deux cerveaux identiques, même ceux de jumeaux. Cela est dû en grande partie au fait que, au cours du développement et de l'établissement de la neuroanatomie, les neurones qui s'activent ensemble se connectent⁸ » Il est difficile de voir exactement en quoi cette critique représente un argument rigoureux à l'encontre de l'ordinateur, puisque les réseaux de neurones peuvent également prendre des jeux de valeurs uniques au cours de leurs apprentissages. Il semblerait qu'Edelman récuse l'idée que le développement neuroanatomique du cerveau puisse être apparenté à *l'intelligence learning* bien connu des programmeurs. Ces derniers en effet définissent un ensemble de règles d'évolution régulées par une horloge logique, là où le cerveau est le fruit d'un processus sélectif comparable à l'immunologie, qui conserve uniquement les réseaux neuroneaux efficaces. À ce titre, dans les yeux d'Edelman, la programmation évolutionnaire, bien qu'inspirée de la biologie, ne capture pas véritablement le fonctionnement de l'organe cérébral.

Le troisième argument d'Edelman réside en ceci que l'être humain ne possède pas suffisamment de gènes pour coder entièrement le fonctionnement du cerveau. Il est certes facile de démontrer que le nombre de connexions synaptiques possibles excède de loin le nombre de combinaisons génétiques, si bien qu'une partie de l'apprentissage n'est pas codé *a priori* dans l'organisme, mais repose sur des phénomènes émergents du contact avec l'environnement. Bien qu'intéressant, il semblerait, une fois encore, que l'argument souffre d'une simplification. Il consiste en effet à tisser une analogie entre le code génétique et le code informatique, de sorte à conclure que, si le cerveau était un ordinateur, il faudrait que son fonctionnement puisse être décrit parfaitement par son code génétique. Cela présuppose implicitement que le fonctionnement d'un ordinateur est totalement décrit par son code. Or en ce qui concerne les réseaux de neurones artificiels, il se trouve que le code ne fait, tout au plus, que définir les règles d'apprentissage à partir desquels les variables clefs peuvent converger vers certaines jeux de valeurs. La disposition finale du réseau de neurones artificiels émerge d'un proces-

8. GERALD M. EDELMAN, *Second Nature : Brain Science and Human Knowledge*, Yale University Press, New Heaven and London, p.21

sus largement indépendant de la volonté initiale du programmeur — si bien qu'on chercherait en vain à le prédire sur base du code informatique initial. Le troisième argument d'Edelman se ramène donc soit au premier — à savoir qu'il n'existe pas de règles logiques dans le cerveau comparable à celles de l'ordinateur — soit il consiste à dire que le code génétique est trop faible pour définir ne serait-ce que l'ensemble des variables synaptiques concernées dans la formation neuroanatomique du cerveau.

Pour comprendre la persistance d'Edelman à s'attaquer à l'existence d'une horloge logique et d'une prédétermination génétique, il est sans doute utile de donner quelques éléments de contexte. Dans les 1970, années pendant lesquelles Edelman développe ses premières théories sur le cerveau, il existait une prévalence importante à l'idée, premièrement, que les gènes co-daient le comportement du cerveau ; et, deuxièmement, qu'ils le faisaient de façon centralisée. Il faut dire que la découverte de l'ADN par James Watson et Francis Crick, encore fort récente, avait conquis les représentations scientifiques dominantes, faisait du gène le fondement d'une explication mécanique intégrale du comportement humaine. Cette « machinerie génétique », comme le désigne Thierry Magnin⁹, devait supposément expliquer le développement de l'intelligence, de la personnalité, etc. Aujourd'hui, la popularité de l'épigénétisme met l'accent sur l'idée que certains gènes, bien que présents dans l'organisme, vont être amenés à s'exprimer — ou non — en fonction des expériences du sujet, de son vécu psychique (stress, l'exercice psychique, la joie, les pratiques spirituelles...) et de l'interaction complexe avec l'environnement. On est passé d'une vision centralisée de la machine humaine à une vision émergeant de l'interaction avec de nombreuses autres composantes.

Un exemple frappant d'épigénétisme, et tout à fait digne de l'introduction d'une authentique série criminelle, peut être trouvée dans le cas de James Fallon. Ce neuroscientifique de l'Université de Californie a cherché toute sa vie à caractériser les composantes cérébrales et génétiques de la psychopathie lorsque, au hasard d'une expérience, il est tombé sur le cliché d'un cerveau présentant en tous points les marqueurs qu'il avait identifié tout au long de ses travaux. Quelle ne fut pas sa surprise de constater qu'il s'agissait de l'IRMf de son propre cerveau ! James Fallon présentait le gène spécifique de la psychopathie, mais le contexte de vie, assez favorable, dans lequel il avait grandi, n'avait pas permis à ce gène de s'exprimer — bien que la nouvelle de cette particularité génétique n'étonna pas particulièrement son entourage¹⁰.

Ces dernières considérations nous conduisent à la question suivante : si la configuration du cerveau n'est pas pleinement décrite par le code génétique,

9. Cette expression, ainsi que d'autres du même acabit, peuvent être trouvées dans ses exposés oraux ou dans son article *La mesure humaine dans le post et le transhumanisme*.

10. Cette anecdote est rapportée par le principal intéressé dans un essai intitulé *The Psychopath Inside : A Neuroscientist's Personal Journey Into the Dark Side of the Brain*.

comment se fait-il, d'après Edelman, que le cerveau se développe ? Cela demande d'exposer sa théorie neurodéveloppementale connue sous le nom de « Théorie de la sélection des groupes neuronaux », ou encore « Le darwinisme neuronal. »

14.3 Théorie de la sélection des groupes neuronaux (TSGN)

Au vu de ce qui a été dit, on comprendra que la théorie de la conscience, d'après Edelman, doit répondre de plusieurs contraintes, en ce compris la diversité et la régularité des réponses apportées par le cerveau, le tout sans présupposer l'existence d'une horloge et de règles logiques assimilant le fonctionnement de l'esprit à celui d'un ordinateur. Dans la mesure, on l'a dit, où le matériel génétique ne suffit pas pour prescrire la configuration du cerveau dans son ensemble, Edelman s'inspire, pour répondre à ce problème, de la théorie qui lui a valu l'obtention du Prix Nobel de médecine avec Rodney Porter : à savoir, la théorie immunologique décrivant le fonctionnement des anticorps.

C'est un fait connu que l'organisme, après avoir été soumis à une certaine souche virale, est capable de s'adapter efficacement à toute autre menace apparentée. Du point de vue immunologiste, cela consiste donc à dire que les anticorps sont capables de reconnaître des molécules étrangères, de mémoriser cette information, et de les neutraliser lorsqu'ils sont confrontés ultérieurement à la même molécule. Une théorie naïve consisterait à penser que les anticorps, de façon analogue à de la pâte à modeler, peuvent envelopper les antigènes qu'ils rencontrent et conserver les contours de ceux-ci même en leur absence. Une population d'anticorps, mise au contact d'une certaine molécule, développe donc mécaniquement la forme requise pour identifier cette molécule précise. Cette façon de voir les choses assimile la mémoire des lymphocytes à une tablette d'argile dans laquelle vient se graver les informations requises. Or il se trouve que cette conception est fausse.

Edelman a développé en 1972 une théorie basée sur un mécanisme darwinien. Au lieu de doter les anticorps d'une capacité de modulation de forme, il est proposé que chaque lymphocyte de l'organisme possède aléatoirement un type distinct d'anticorps à sa surface, capable de reconnaître et de se lier à un antigène donné. Comme il existe près de cent milliard de lymphocytes dans l'organisme, le corps dispose donc d'un large stock d'anticorps diversifiés. Dès lors qu'une molécule étrangère est introduite dans le sang, les collisions aléatoires de cette molécule avec les lymphocytes va entraîner à un moment donné la rencontre d'un anticorps capable de se lier à cette molécule. Il n'est donc pas besoin de supposer un changement de forme. Le lymphocyte porteur de l'anticorps approprié va se dupliquer de sorte à proliférer. Si la même molécule est introduite dans l'organisme ultérieurement, il

y aura donc une proportion beaucoup plus importante d'antigènes capables de se lier à elle, de l'identifier et de la neutraliser. Ce mécanisme explique la capacité d'apprentissage immunologique à la base, notamment, des vaccins.

Le mécanisme de sélection repose sur trois ingrédients. D'une part, ce qu'on appelle un générateur de diversité, responsable de la population initiale très variée d'anticorps. Ensuite, il faut qu'il y ait une interaction avec l'environnement, de sorte que chacun de ces anticorps soient susceptibles, ou non, de se lier à une molécule étrangère. Pour finir, il faut un mécanisme expliquant la duplication en grand nombre des anticorps ayant trouvé leur cible. D'un point de vue général, ces trois ingrédients sont appelés la diversification, la stimulation et la prolifération. La question qu'il reste à adresser maintenant consiste à se demander si ces trois composants peuvent être trouvés dans la formation des cellules nerveuses ?

Edelman répond par la positive. Il défend que la diversification des cellules nerveuses est rendue possible en raison de la forte variation microscopique observée dans le développement neuroanatomique. Tout d'abord, il convient de noter qu'il existe près de deux cents types de neurones distincts, qui varient par leur stimulabilité et leur conductivité. La stimulabilité représente le potentiel d'entrée nécessaire pour que le neurone provoque une impulsion neuronale, tandis que la conductivité dépend de la forme des neurites, et particulièrement la longueur de l'axone et le nombre de dendrites (voir figure 14.1). À titre d'exemple, un neurone pyramidal du cortex moteur présente un très long axone assurant une communication avec les régions du système nerveux. À l'inverse, le neurone de Golgi de type II présente un axone très court et assez peu de terminaisons nerveuses. On le trouve dans le cortex cérébral, le cortex cérébelleux et dans la rétine.

Non seulement il existe un grand nombre de neurones différents, mais les liaisons entre ceux-ci sont, à leur tour, susceptibles d'une grande diversité évolutive. Une représentation simplifiée du cerveau consiste à dire que la population neuronale, depuis son développement fœtal jusqu'à un stade plus avancée de son évolution, présente des excitations aléatoires. Or en vertu de la règle de Hebb, deux neurones qui s'activent ensemble se lient conjointement. Plus rigoureusement, on dira de deux groupes de neurones A et B qu'ils sont connectés si le potentiel d'action, ou la décharge, nécessaire dans l'un des groupes pour faire réagir le second est plus faible que pour faire réagir d'autres groupes de neurones. Tout se passe donc comme si des chemins privilégiés se créaient entre les cellules du cerveau au fur et à mesure des expériences ; que la résistance au passage de l'information se réduisait avec l'habituation — exactement comme l'herbe d'un sentier se raréfie sous la semelle des promeneurs.

Bien sûr, l'évolution du cerveau n'est pas purement et simplement le fruit du hasard. Les excitations neuronales, et il s'agit là du deuxième ingrédient de la théorie darwinienne d'Edelman, vont répondre aux contraintes de l'environnement. La connexion entre deux neurones, ou groupes de neu-

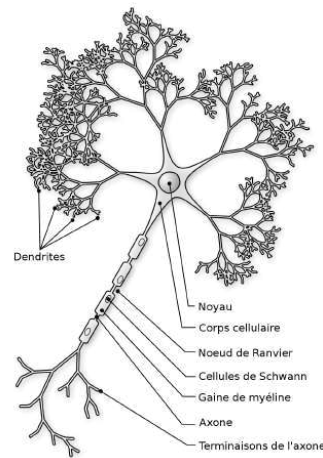


FIGURE 14.1 – De façon schématique, un neurone est composé essentiellement d'un noyau, de dendrites et d'un long axone. Les dendrites reçoivent les signaux afférents. Si leur sommation conduit à l'activation du neurone, un potentiel électrique est communiqué le long de l'axone, via les cellules de Schwann, jusqu'à d'autres neurones. La réception d'une impulsion neuronale peut être facilitée ou inhibée par la présence de neurotransmetteurs (source de l'image : Wikipédia, article francophone sur le neurone.).

rones, est d'autant plus encouragée que les réponses physiologiques résultant de cette corrélation d'activité sont adéquates pour répondre aux besoins de l'organisme et à sa survie dans la nature. Un exemple typique est le fonctionnement de la mémoire chez l'aplysie (un escargot marin géant), étudié par Kandel^[11]. Lorsque cet animal reçoit un choc sur une partie du corps telle que la queue, il reconnaît le caractère désagréable du stimulus et apprend à augmenter son réflexe de défense vis-à-vis d'un grand nombre de stimulus ultérieurs, ce qui se manifeste par la contraction des muscles, une attitude de retrait, une plus grande réactivité, etc. « Un choc unique donne naissance à une mémoire qui ne dure que quelques minutes ; cette mémoire à court terme ne requiert pas la synthèse de nouvelles protéines. Au contraire, quatre ou cinq chocs espacés dans le temps au niveau du pied donnent naissance à une mémoire qui va durer plusieurs jours^[12] » Cela résulte d'une connexion indirecte établie entre les neurones sensitifs et les neurones moteurs, s'opérant par la médiation d'un petit groupe d'interneurones excitateurs et inhibiteurs.

Il reste à expliquer ce qui assure la prolifération des neurones, ou des connexions neuronales, qui ont été sélectionnées au contact de l'environnement. Dans le cadre de l'apprentissage de l'aplysie, les travaux de Kandel

11. ERIC R. KANDEL, *La biologie moléculaire de la mémoire : un dialogue entre gènes et synapses*, M/S : médecine sciences, 19(5), 625–633.

12. *Ibidem*, p.627

ont montré que la mémoire du traumatisme est rendu possible via un mécanisme impliquant la création d'une certaine protéine (appelée CREB-1) au sein des cellules nerveuses, et impliquant l'ARN messager. La création de la protéine est d'ailleurs ce qui différencie la mémoire à court terme de la mémoire à long terme. Edelman va cependant pointer un autre phénomène : la réentrée. La réentrée est le passage continu de signaux d'une région du cerveau (ou carte) à une autre, et leur retour, à travers les fibres massivement parallèles (axones) dont on sait qu'elles sont omniprésentes dans les cerveaux supérieurs. Ce mécanisme permet d'expliquer pourquoi on observe l'activation concomitante de différentes parties du cerveau en l'absence d'un administrateur central, ou de tout organe chargé de coder de façon top down le fonctionnement du cerveau. C'est donc à ses yeux l'argument central récusant l'assimilation du fonctionnement cérébral à celui d'un ordinateur. Pour illustrer ce phénomène de réentrance, le lauréat du prix Nobel ne manque pas de produire une analogie imaginative :

Pour imaginer comment fonctionne la réentrée, considérez un quatuor à cordes hypothétique composé de musiciens talentueux. Chacun joue son air à un rythme différent. Reliez maintenant le corps de tous les musiciens avec de tout petits fils (nombreux sur toutes les parties du corps). Lorsque chaque musicien bouge, il envoie inconsciemment des ondes de mouvement aux autres. Au bout d'un court moment, le rythme et dans une certaine mesure les mélodies deviendront plus cohérents. Cette dynamique se poursuivra, donnant une nouvelle sortie cohérente. C'est quelque chose de ce genre qui se produit dans le jazz improvisé, sans fils, bien sûr¹³!

Le mécanisme de réentrée explique d'après Edelman la possibilité de former des idées cohérentes en l'absence d'un organe de régulation. On se souvient que, pour Descartes, la glande pinéale occupait un rôle central en ceci qu'elle permettait de synthétiser les informations dupliquées issues des sens (les deux yeux, les deux oreilles, les deux narines...) de façon à donner naissance à une idée homogène, claire et distincte, de l'objet perçu. Or d'après le modèle théorique d'Edelman, l'hypothèse de la glande pinéale, ou de tout autre appendice central, est devenue superflue. C'est au contraire de la coordination concomitante des aires cérébrales liées à l'orientation d'un stimulus (aire V1), à la couleur (V4) et à son mouvement (aire V5) que va naître le percept unifié de la vision.

Notons que la réentrée a reçu un bel avenir dans les théories neuroscientifiques contemporaines, puisqu'il se trouve non seulement au cœur de la théorie de Giulio Tononi, qu'on verra dans le prochain chapitre, mais également dans les modèles de la conscience d'accès de Dehaene, Changeux ou

13. GERALD M. EDELMAN, *Seconde Nature : La Science du Cerveau et la Connaissance*, Odile Jacob, p. 44

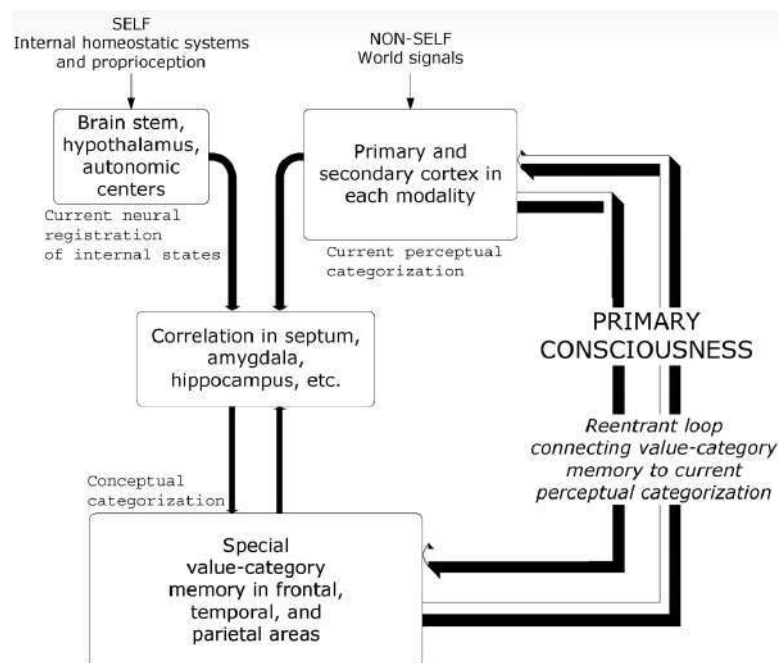


FIGURE 14.2 – Exemple de système de réentrance proposée par Gerald Edelman (source : GERALD M. EDELMAN, *Naturalizing consciousness : A theoretical framework*, PNAS, avril 2003).

Naccache.

14.4 Théorie edelmanienne de la conscience

En quoi la théorie neuronale darwinienne qui vient d'être exposée peut-elle servir de support à une théorie scientifique ou philosophique de la conscience ? Le cas échéant, qu'est-ce qu'une telle théorie implique concernant des facultés liées à l'exercice de la conscience, telle que la raison, l'imagination, l'intuition ou la créativité ? Avant de s'attaquer à de telles problématiques, qui sont au cœur de sa démarche, Edelman précise d'emblée que la conscience phénoménologique, définie comme le fait d'être, à un instant donné, traversé d'un vécu qualitatif et subjectif, ne doit pas être confondue avec la conscience de soi, ou la conscience réflexive, qui consiste à prendre pour objet de conscience une certaine représentation de soi en tant que sujet pensant. C'est la même distinction qui existe entre le « Je pense, je suis » en tant que fait irréductible de conscience et le fait même de penser : « Je pense donc je suis ».

Dans un ouvrage écrit de concert avec Tononi¹⁴, Edelman défend l'idée que la conscience est apparue lorsque le système thalamocortical s'est considérablement agrandi et que le nombre de noyaux thalamiques spécifiques a augmenté sous l'élargissement du cortex cérébral. Ces conditions favorisent en effet l'activité réentrante entre les aires corticales et le thalamus. Il est postulé que cela remonte à la période marquant la transition entre les reptiles et les oiseaux, puis les mammifères, il y a près de deux cent cinquante millions d'années. Il est difficile de concevoir comment cet énoncé peut avoir une valeur scientifique, même si l'argument anatomique reste justifié par les conceptions exposées au chapitre précédent. Du reste, cela pose la question de savoir ce que ressentait les dinosaures ou les animaux issus de l'explosion cambrienne ? Peut-être faut-il distinguer la sentience de la conscience proprement dite ? Peut-être que la théorie d'Edelman, tout en expliquant la conscience de l'être humain, n'exclut pas que des vécus de conscience, quoique différents de notre espèce, puissent se produire d'après d'autres mécanismes physiologiques que ceux spécifiques à notre cerveau ?

Toujours est-il qu'Edelman est plus optimiste que son homologue Julian Jaynes, qui faisait remonter la conscience, quant à lui, à une époque succédant de près à l'apparition du langage et des métaphores¹⁵, cherchant

14. GERALD H. EDELMAN, G. TONONI, *Comment la matière devient conscience ?*, Odile Jacob, 1er mai 2000.

15. Dans son ouvrage intitulé *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Julian Jaynes fait remonter l'émergence de la conscience de soi à une époque avoisinant le deuxième millénaire avant Jésus-Christ. Il stipule que, avant cette date, dominait un esprit bicaméral, inspiré des deux hémisphères cérébraux : une partie de cet esprit était composée de voix et l'autre partie écoutait celles-ci. La notion de soi comme d'une entité unitaire et autonome était fragmentées entre ces deux instances.

même dans Homère l'attestation d'une vie mentale bien différente de celle que nous connaissons aujourd'hui.

D'après Edelman, les animaux disposant d'un cerveau avec une forte capacité de réentrance seraient en mesure de catégoriser les informations sensorielles de façon raffinée et d'associer les valeurs synaptiques correspondant à chacune de ces catégories à certaines zones de la mémoire. Lorsqu'une scène est perçue à un instant donné, des catégories seraient extraites par le cerveau thalamocortical, tandis que le système de la mémoire quant à lieu serait médiatisé par des régions corticales plus antérieures, comme le cortex frontal et pariétal. De cette façon, il est en mesure de structurer l'information incidence de façon à former des percepts unifiés : plutôt que de voir le monde comme une rapsodie de données sensibles, se bousculant dans l'esprit humain avec la violence de la marée contre la digue, les contours d'objets déterminés prennent vie au milieu des flots. Il serait intéressant de se demander à quoi ressemblerait le monde sensible sans la présence de ces percepts ? Serait-ce une mer d'huile, homogène en tout point de l'espace, une espèce d'être parméniidien sans mouvement, sans aspérités ? Ou serait-ce au contraire une tempête digne du chaos primordial des grandes mythologies ? Cette question sera traitée dans le chapitre...

Cette conscience qui a été décrite pourrait être qualifiée de primaire, car elle ne porte pour ainsi dire que sur les états présents. D'un point de vue neuroscientifique, Edelman suggère que la conscience de soi apparaisse en ajoutant la possibilité, pour le cerveau, de porter sur ses propres états.

14.5 Chiasmes métaphysiques et principes épistémiques

Bien qu'Edelman nie toute forme du dualisme, ou peut-être à cause de cela, sa théorie, pour recevoir une signification complète, s'appuie indirectement sur une psychologie du sens commun. Lorsqu'il écrit, pour décrire le phénomène de la mémoire, que « le cerveau se parle à lui-même¹⁶ », il tisse une correspondance implicite entre un phénomène physiologique, impliquant de nombreux mécanismes biologiques darwiniens, et une expérience vécue par un sujet humain. De manière générale, Edelman donne ainsi une savante description du fonctionnement du cerveau en tant qu'organisme, mais il n'adresse pas véritablement la question difficile de la conscience : à savoir pourquoi se fait-il que l'on sente quelque chose plutôt que rien ? Pourquoi se fait-il que de la synergie de tous ces mécanismes biochimiques naissent, pour ainsi dire, un phénomène conscient ?

Edelman ne doit pas pour autant être taxé de naïveté philosophique — épithète d'autant plus difficile à attribuer qu'elle constitue, pour l'académique contemporain, une sorte de cauchemar ambulant. Bien qu'il s'attaque

16. Edelman, *op. cit.*, p. 47

au dualisme et tente de bâtir une théorie de la conscience sans recourir explicitement à l'introspection, Edelman concède que la théorie de la conscience, pour acquérir une signification, doit reposer sur un dualisme. Ce qu'il nie simplement, c'est que l'esprit en tant que tel puisse avoir une action causale sur la matière. Ou autrement dit, il ne veut pas conférer à des entités abstraites, telles que la volonté, le soi ou la mémoire, une signification réelle au-delà des idées que l'être humain s'en fait. En cela, il s'oppose certes à la glande pinéale de Descartes, mue par la seule force de l'âme, mais il se rapproche aussi de la position que je cherche à défendre, au sein de laquelle les modalités du corps et de l'esprit n'ont pas d'action directe l'une sur l'autre : elles se rapportent à une même réalité fondamentale.

Pour comprendre que c'est l'activité réentrante et sélectionniste des groupes de neurones faisant partie du noyau qui entraîne la conscience phénoménale, il est nécessaire d'évoquer le dualisme. Même si la conscience est un processus en lui-même dépourvu de pouvoir causal, elle n'en est pas moins entraînée par l'activité complexe et le pouvoir causal des groupes de neurones qui constituent le noyau réentrant. De plus, depuis les tout premiers temps du développement, les signaux allant du corps au cerveau et de ce dernier à lui-même forment les bases de l'émergence d'un soi^[17].

Edelman parle de pouvoir causal, ce qui laisse à penser que, pour lui, les activités cérébrales puissent causer des états mentaux, tels que la perception ou la conscience de soi. Il précise cependant (a) que ces états mentaux, ou qualia, ne sont pas des choses du cerveau, mais bel et bien des processus physiologiques liés au traitement de l'information neuronale et (b) que la notion de cause n'est pas tout à fait exacte, et que l'on devrait plutôt parler de support à l'apparition du phénomène : « Les hypothèses selon lesquelles c'est l'action du noyau thalamocortical qui cause la conscience se heurtent à une difficulté. Puisque les causes précèdent les effets, ces hypothèses impliquent l'existence d'un écart temporel entre processus incommensurables. Au contraire, l'action neurale dans le noyau entraîne la conscience, tout comme le spectre de l'hémoglobine du sang est entraîné par la structure mécanique et quantitative de cette molécule^[18]. »

Dans son livre *Plus vaste est le ciel : Une nouvelle théorie générale du cerveau*, Edelman suggère même de dresser un parallélisme psycho-physique, avec d'un côté les processus mentaux tels que nous les vivons, peuplés de souvenirs, d'expériences, de représentations — ou, en un mot, de qualia — et de l'autre côté les processus cérébraux composés de neurones, de myéline, d'axones et autres substances cérébelleuses^[19]. Bien qu'il n'y ait pas de cau-

17. EDELMAN, *Seconde Nature : La Science du cerveau et la connaissance*, Odile Jacob, p. 53.

18. *Ibidem*, p. 57.

19. Se référer au chapitre *La transformation phénoménal* de l'ouvrage cité ci-dessus

salité directe de l'esprit sur la matière, ou de la matière sur l'esprit, Edelman défend que les causes identifiables entre les neurones correspondant, pour ainsi dire analogiquement, avec les causes entre phénomènes mentaux. Cette assertion est très proche du monisme structurel que je défends dans cette thèse, mais elle présente deux différences majeures.

D'une part, elle continue de voir dans l'esprit et la matière des substances à part entière (le mental et la matière), au risque de contredire l'idée d'Edelman d'après laquelle la conscience n'est pas une chose, mais un processus. D'autre part, l'esprit est envisagé en des termes assez naïfs. Il est supposé que le Soi, le cœur de la conscience, est cause de la production de pensées et de sensations. Il y a, dans un cas comme dans l'autre, réification de la conscience. Cette dernière est vue comme une entité objective agissant dans le monde. À rebours de cette attitude, je défends la thèse d'après laquelle la conscience est un phénomène qui se vit mais qui ne peut pas être représenté comme un objet au sein du monde mental. Loin d'être au cœur de la vie de l'esprit, la conscience est un horizon, une frontière qui ne se laisse pas appréhender directement, mais dont le concept est toutefois nécessaire, et même, signifiant.

L'idée du parallélisme est séduisante, mais il faudrait idéalement tisser une analogie plus poussée entre l'expérience de la mémoire, de la créativité, de l'imagination, telles que nous les vivons, et les principes au fondement du fonctionnement neuronal, afin de montrer qu'il s'agit, in fine, de la même structure processuelle. Ceci pourrait être réalisé en élaborant, en parallèle de la théorie du cerveau, une théorie de la conscience vécue, et en montrant que cette théorie répond aux mêmes principes épistémologiques que la théorie de la sélection des groupes neuronaux. Ce parallélisme ne présupposerait cependant pas deux substances distinctes, dotées de leur pouvoirs causaux, mais l'existence d'une même réalité psycho-physique pouvant être vue d'après deux perspectives distinctes.

Une telle approche sera envisagée dans la troisième partie de cette thèse. Pour le moment, j'aimerais récapituler ici l'ensemble des grandes lignes théoriques directrices de la théorie d'Edelman²⁰ :

- **Énactivisme** : Edelman voit la conscience comme le fruit de trois ingrédients : la présence d'un cerveau, le fait que ce cerveau soit enveloppé dans un corps et l'interaction de ce corps avec l'environnement. À ce titre, il s'oppose à une vision computationnelle, qui assimile le fonctionnement de l'esprit à celui de l'ordinateur. Il n'existe pas dans l'organe cérébral d'horloge centralisée, permettant, à chaque unité de temps, de réaliser un certain nombre d'opérations logiques. Edelman

20. Pour une discussion plus détaillée de l'épistémologie d'Edelman, je réfère le lecteur à l'article de Edmond Ortigues intitulé *Edelman et la conscience humaine*, Éditions de Minuit, dans Critique 2005-12 (Num. 703), 2005-12 (Num. 703), pages 917 à 926.

ne nie toutefois pas le rôle crucial du cerveau dans le phénomène de la conscience.

- **Évolutionisme** : Pour cette même raison que le cerveau n'est pas un ordinateur, son fonctionnement n'est pas préprogrammé par le matériel génétique de l'individu. Les réseaux neuronaux, tout au contraire, sont sélectionnés au contact de l'environnement par un processus darwinien. Celui-ci est rendu possible en raison de trois facteurs : (1) une grande diversité neuronale préalable ; (2) le renforcement des liaisons synaptiques grâce à la règle de Hebb, qui stipule que deux neurones qui s'excitent ensemble se lient ensemble ; (3) la présence d'une activité réentrante, qui garantit la forte intégration des neurones dans un environnement plus large. « La réentrée, dit Edelman, c'est l'échange récursif permanent des signaux parallèles entre les aires cérébrales qui sert à coordonner l'activité des différentes zones dans l'espace et dans le temps. Par opposition au feed-back, la réentrée n'est pas une transmission séquentielle d'un signal d'erreur dans une boucle simple. Elle implique simultanément beaucoup de voies réciproques parallèles et n'a pas de fonction d'erreur prescrite qui lui soit attachée²¹ »
- **Émergentisme processuel** : Edelman reprend à William James cette idée que la conscience ne saurait être vue purement comme une chose, mais bien comme un processus. Contrairement à un atomisme logique ou psychologique, il ne s'agit donc pas de réduire la conscience à tout autant d'aptitudes cognitives ou psychologiques (le réflexe, la sensation, la vision, le langage, la mémoire...) puis d'associer ces aptitudes à des parties du cerveau (le cortex sensorimoteur, l'aire de Broca, l'aire de Wernicke, l'hippocampe...) La conscience au contraire est un processus au sein d'un système complexe. « Un système complexe, précise Edelman, est un système qui consiste en diverses petites parties, chacune pouvant être isolée d'un point de vue fonctionnel. Comme ces parties hétérogènes interagissent selon des combinaisons variées, on note une tendance à donner naissance à des propriétés du système qui sont davantage intégrées²² »
- **Causalité émergente** : La position d'Edelman en ce qui concerne la causalité est tâtonnante, mais il est clair qu'il s'oppose à l'idée que la matière agisse sur le mental, ou qu'inversement le mental agisse sur la matière. Il semble voir dans les substrats neuronaux un support à l'apparition des phénomènes mentaux. À ce titre, il semble s'inscrire

21. GERALD M. EDELMAN, *Plus vaste est le ciel : Une nouvelle théorie générale du cerveau*, Odile Jacob, Paris, p. 57

22. *Ibidem*, p. 87

dans une théorie émergentiste qui n'est pas sans rappeler l'idée de survenance défendue par Jaegwon Kim²³. À certains endroits, il préfère toutefois parler de parallélisme psycho-physique qui s'apparente davantage à une conception leibnizienne.

23. Voir notamment son article *Making Sense of Emergence*, ainsi que son livre *La Survenance et l'esprit*.

Gerald Edelman est connu pour avoir transposé à l'étude de la conscience sa théorie sur la sélection des immunoglobulines, qui lui a valu le prix Nobel en 1972.

La théorie sur la sélection des immunoglobulines stipulait que les anticorps sont originellement contenus dans le sang sous des formes variées. En présence d'un agent pathogène, ceux ayant la forme requise pour contrer la menace vont se multiplier rapidement, entraînant *de facto* une plus grande résistance de l'organisme face à des infections ultérieures.

En conservant l'esprit darwinien de ce mécanisme, Edelman va faire trois hypothèses sur le cerveau humain :

1. La population des neurones présente, initialement, des formes et des connexions diverses.
2. En présence d'une stimulation de l'environnement, certains neurones vont être excités simultanément, d'autres non. Les neurones qui se sont excités en même temps vont se lier statistiquement (loi de Hebb).
3. Les groupements de neurones liés vont présenter une forte activité de « réentrance » : les signaux entrants dans un système cérébral vont passer dans d'autres systèmes avant d'être réinjectés dans le système initial, et ainsi récursivement.

La conscience, d'après Edelman, ne doit pas être conçue comme un produit des neurones eux-mêmes, considérés en tant qu'objets physiques, mais comme le processus dont ces neurones sont le support. Elle émerge du traitement circulaire de l'information.

Edelman défend ainsi une vision non-réductionniste de la conscience, teintée d'énactivisme. Il suppose néanmoins une causation implicite entre les phénomènes de conscience vécus et les activités du cerveau. À ce titre, il est permis de suspecter, dans sa pensée, la présence d'un chiasme métaphysique externaliste.

Chapitre 15

La Théorie Intégrée de l'Information

La conscience n'est pas un objet ; c'est un processus.

Giulio Tononi

L'histoire de la pensée se présente comme une grande toile plane, que des idées plus ou moins lourdes en significations viennent déformer tour à tour. Chaque contribution infléchit insensiblement le champ du savoir, créant tantôt des terrains favorables, tantôt des gouffres sans issue. À cet égard, la théorie de Gerald Edelman ne fait pas exception. D'une part, elle ouvre un champ d'interrogation nouveau, où la conscience est conçue comme un phénomène évolutionniste émergeant, reposant sur la haute intégration des réseaux neuronaux, et non plus comme une simple propriété cognitive, localisable dans une partie du cerveau à la façon dont on associait la production du langage à l'aire de Broca, la compréhension du langage à l'aire de Wernicke ou la mémoire à l'hippocampe.

Mais ce terrain prometteur, plus proche d'une jachère que d'un champ à la saison des moissons, manque cruellement de repères. Bien qu'Edelman insiste sur le rôle du système thalamocortical, qui sert de relais aux informations sensori-motrices entre le cortex et le thalamus¹, il ne donne pas de critère définitif permettant d'identifier les structures d'ordre supérieures impliquées dans le phénomène de la conscience. Les patients cérébrolésés, dont seule une faible portion du cortex demeure intacte, sont-ils conscients ? Les nouveau-nés sont-ils conscients ? Les animaux qui affichent des comportements complexes, mais dont le cerveau est très différent de celui des

1. Dans son ouvrage technique intitulé *The Remembered Present : A biological Theory of Consciousness*, Gerald Edelman situe dans le système limbique et le tronc cérébral l'origine du soi biologique, qui donne à l'organisme ses objectifs et ses valeurs, tandis qu'il voit le système thalamocortical une capacité à opérer des comparaisons continues entre le soi et le non soi.

humains, sont-ils conscients ? Les machines intelligentes peuvent-elles être conscientes^[2] ?

Edelman s'interroge très tôt sur l'existence d'un substrat physique à la conscience. En 1999, il publie un article avec Giulio Tononi intitulé *Consciousness and Complexity*, qui jette les bases de ce qui deviendra plus tard la théorie intégrée de l'information. Ils prennent un double parti : celui d'aller au-delà de la position holistique de William James, d'après laquelle le cerveau en entier est responsable de la conscience^[3], et au-delà également de la position localisationniste de certains neuroscientifiques, qui attribuent à des neurones, ou groupes de neurones précis, le pouvoir de créer des phénomènes conscients. Cette dernière approche, disent-ils en empruntant les mots de Gilbert Ryle, représente une authentique « erreur de catégorie »^[4], puisqu'ils donnent un pouvoir proprement mental et sensitif à des unités matérielles. Pour éviter cet écueil, quelle voie intermédiaire reste-t-il à emprunter ?

Les auteurs suggèrent de s'intéresser, non pas à des groupes de neurones particuliers, ni même à des régions cérébrales, mais à des types de fonctionnement neuronaux. En admettant que la conscience soit avant tout un processus de traitement de l'information, et non un organe ou un objet parmi d'autres, leur stratégie consiste à isoler le mécanisme neuronal le plus susceptible d'épouser les propriétés de ce processus, puis seulement ensuite, dans un second temps, de chercher où dans le cerveau ce mécanisme peut être observé.

Cela conduit à une nuance de taille avec l'approche purement localisationniste : premièrement, il n'est pas nécessaire que ce soit, à tout moment, les mêmes régions du cerveau qui implémentent les mécanismes neuronaux de conscience. Il est au contraire tout à fait possible que, pendant le sommeil, certaines parties du cerveau différentes de celles de l'éveil prennent le relais des mécanismes conscients. Par ailleurs, il se peut que des variations existent en fonction des types de tâches réalisées (visuelles, auditives, souvenirs...), des individus ou de la présence de lésions. Enfin, il est admissible, en théorie, que ces processus soient implémentés par d'autres substrats que ceux auxquels le cerveau humain nous a familiarisés — que l'on pense aux cerveaux des insectes ou aux machines de silicium.

Pour identifier les processus neuronaux en question, Edelman et Tononi adoptent une méthodologie inversée. Au lieu de partir du cerveau et d'y chercher, à tâtons, la présence d'une activité consciente, ils prennent l'expérience phénoménologique pour point de départ, dans l'espoir de découvrir des critères qui, seulement ensuite, serviront d'indice pour repérer les processus neuronaux adéquats. La question de recherche initiale devient donc :

2. Ces questions sont celles proposées en exemple par Tononi et ses collègues dans leur article d'opinion intitulé *Integrated Information Theory : From Consciousness to its physical substrate*, *Nature Review Neurosciences*, Mai 2016.

3. JAMES WILLIAMS, *The Principles of Psychology*, Holt, New York, 1890.

4. RYLE G., *The Concept of Mind*, Hutchinson, London, 1949.

qu'est-ce que cela veut dire que de vivre une expérience de conscience ?

Edelman et Tononi répondent en isolant deux composantes clés dans les processus neuronaux : l'*intégration* et la *différenciation*. L'intégration est le fait que les contenus mentaux se présentent à la conscience de façon unifiée, c'est-à-dire, comme une totalité insécable. Il s'agit d'une scène, mobilisant certes différents sens internes et externes, mais au sein de laquelle tout est donné d'un seul bloc. À ce titre, l'individu est contraint, pour ainsi dire, de voir le monde depuis une perspective privée, intime.

La *différenciation*, quant à elle, consiste en ceci que, bien que tout acte de conscience se présente comme un tout unifié, l'esprit est capable de décomposer ce tout en éléments distincts, procédant à cet effet en de multiples comparaisons. L'intervalle de temps requis pour décomposer une scène est extrêmement court. Pour la seule discrimination visuelle, une fraction de seconde suffit à distinguer, au sein d'une image, une pléthore d'objets distincts. Il y a donc présence d'un processus opérant une différenciation entre les informations, de sorte à isoler des concepts ou de schémas au sein de l'ensemble.

Une fois que les propriétés phénoménales ont été identifiées, il reste à formuler ces dernières dans un langage quantitatif permettant de déterminer si, oui ou non, un processus neuronal peut être qualifié de conscient. Pour ce faire, Tononi et Edelman recourent à l'entropie, une grandeur développée par Shannon pour mesurer la quantité d'information entre plusieurs variables statistiques⁵.

Plus précisément, ils estiment que chaque neurone se comporte comme une variable, *a priori* aléatoire, susceptible de prendre la valeur 1 (excitation) ou 0 (repos), et exerçant une influence probabiliste sur l'activité ou le repos des autres neurones. La question de savoir si un ensemble de neurones forment une unité fortement intégrée revient dès lors à se demander s'il présente une forte indépendance statistique par rapport au reste du système neuronal. Cela peut être déterminé mathématiquement grâce à une fonction entropique appelée l'information mutuelle, qui mesure la perte d'incertitude acquise sur un ensemble de variables aléatoires en raison des informations détenues à propos d'un autre ensemble de variables aléatoires⁶.

5. Pour plus de détails sur les outils techniques, se référer à l'annexe 6.

6. Le problème consistant à isoler les composantes statistiquement indépendantes d'un ensemble d'information remonte au *cocktail party problem* : lorsqu'un ensemble d'individus sont rassemblés dans un même espace et discutent en petit groupe, comment se fait-il que le cerveau soit capable d'isoler, du fond ambiant, les informations pertinentes à la conversation ? L'explication neurologique vient de ce que l'être humain est doté de deux récepteurs (ses deux oreilles). Il est possible de croiser les informations sonores issues de chacune de ces oreilles de façon à isoler une conversation du bruit ambiant. Cette dernière est en effet statistiquement indépendante du bruit environnant. Les mesures mathématiques utilisées pour modéliser ce genre de problème utilisent tantôt l'analyse en composantes indépendantes (basée sur l'information mutuelle), tantôt l'analyse en composantes principales ou, simplement, la divergence de Kulbach-Leibler.

À partir de ces travaux initiaux, réalisés de concert avec Edelman, Tononi va poursuivre plus avant le perfectionnement de l'approche théorique au travers de plusieurs versions successives de la théorie intégrée de l'information 1.0, 2.0, 3.0 et 4.0. Il dote progressivement sa théorie d'un fondement épistémologique, et ce à l'aide de quatre ingrédients : (1) tout d'abord, il définit un ensemble de principes méthodologiques, basés sur une forme de réductionnisme opérationnel ; (2) il donne les axiomes visant à décrire le plus simplement possible ce que représente une expérience de conscience authentique et (3) il fait correspondre ces axiomes à des postulats physicalistes, dotés chacun d'un outil de mesure quantitatif. Son idée phare est d'établir une analogie entre le fonctionnement phénoménologique de l'expérience consciente et les mécanismes physiques cachés derrière certains processus neuronaux. Cela l'amène à discuter certaines conséquences de sa théorie concernant l'ontologie^[7], le libre arbitre^[8] ou encore l'information^[9].

Dans ce chapitre, je vais donner un succédané de la théorie de Tononi, dans sa dernière version en date (la version 4.0, sortie en décembre 2022), après quoi j'entreprendrai une mise en perspective philosophique.

15.1 Stratégie de l'équivalence structurelle

Tononi entend fonder une théorie de la conscience qui, à l'instar de Gerald Edelman, prenne au sérieux l'expérience phénoménologique, entendue comme le fait irréductible de sentir quelque chose plutôt que rien. Dans l'introduction de son article intitulé *Integrated information theory (IIT) 4.0 : Formulating the properties of phenomenal existence in physical terms*, il écrit : « Être conscient — avoir une expérience — signifie "que ça fait quelque chose que d'être conscient" : ça fait quelque chose que de contempler le ciel, d'entendre le grondement de l'océan, de rêver du visage d'un ami, d'imaginer le flot d'une mélodie, de considérer un choix ou de réfléchir sur sa propre expérience de conscience. »

Tononi estime ainsi que, en l'absence d'un lien explicatif, la simple description des propriétés neurologiques du cerveau ne suffit pas à comprendre pleinement le phénomène de la conscience. Il ne souhaite cependant pas instaurer un dualisme cartésien, qui opposerait d'une part la modalité de la conscience et, d'autre part, celle du corps. Conformément à l'esprit scientifique, il veut rendre compte de l'expérience consciente en des termes physicalistes, mais de manière telle qu'un éclairage, ou un sens, puisse être donné à l'expérience de conscience.

Pour ce faire, il commence par définir un ensemble d'axiomes décrivant

7. TONONI G., *On being* ; à paraître.

8. TONONI G. et al., *Only what exists can cause : An intrinsic view of free will*, arXiv, Juin 2022.

9. TONONI G. et al., *System Integrated Information*, Georg Northoff, 11 february 2023.

les propriétés élémentaires de l'expérience phénoménologique, après quoi il pose un ensemble de postulats concernant le fonctionnement physiologique de l'esprit proprement dit, envisagé comme une certaine façon de traiter l'information. L'objet de la démarche consiste dès lors à montrer que les axiomes de l'expérience phénoménologique coïncident, pour ainsi dire, avec les postulats physicalistes ; on peut rendre raison des premiers au regard des seconds et, inversement, les seconds s'appuient sur les premiers. Cette approche, que l'on pourrait qualifier d'*analogisme structurel* — et que Tononi appelle quant à lui une *identité explicative* — s'appuie sur cinq principes épistémologiques :

1. **Inférence à la meilleure explication** : les postulats de l'esprit sont des hypothèses choisies de façon à rendre compte, le plus efficacement possible, non seulement des phénomènes neurologiques connus, mais aussi des axiomes de l'expérience phénoménologique. La théorie qui en résulte doit idéalement embrasser un grand nombre de faits (amplitude), de façon unifiée (synthèse), en expliquant les faits de façon précise (spécificité), sans contradiction interne (consistance systémique), être sujet à des prédictions expérimentales (validité scientifique), tout en restant plus simple que les théories concurrentielles (simplicité)¹⁰
2. **Réalisme** : Il est fait l'hypothèse que quelque chose existe et persiste en dehors de l'expérience de conscience. L'expérience du monde extérieur sert de condition pour l'élaboration des postulats physicalistes. Tononi défend en outre que l'hypothèse du réalisme, contrairement au solipsisme (qui nie la réalité extérieure), est la seule capable de donner lieu à des explications et des prédictions sur le monde. Il est possible que cette dernière assertion ne dissimule une certaine naïveté métaphysique. On pourrait imaginer un idéalisme au sein duquel la réalité serait toute entière contenue dans l'esprit, bien que cet esprit lui-même serait régi par des lois stables et universelles. L'hypothèse du réalisme n'est cependant pas surprenante dans le cadre d'une démarche scientifique, mais je me demande dans quelle mesure elle s'avère indispensable à la théorie de Tononi ?

10. Il faut prendre garde que l'inférence à la meilleure explication (IME) soit souvent confondue avec l'abduction. La première consiste à tenir pour vraie l'hypothèse qui explique le mieux les hypothèses, tandis que la seconde postule, selon Pierce, à faire l'hypothèse des causes inobservables tels que la présence de particules invisibles ou l'existence d'un événement historique révolu. Si l'IME première s'en tient à des observations, et s'avère ainsi plus sûre, l'abduction en revanche s'appuie sur l'imagination. Elle est moins sûre, mais elle peut s'avérer plus féconde (uberté). Pour de plus amples détails concernant cette distinction, se référer à FRÉDÉRIC ROUDAUT, *Comment on invente les hypothèses : Peirce et la théorie de l'abduction*, dans Cahiers philosophiques 2017-3 (Num 150), pages 45 à 65.

3. **Physicalisme opérationnel** : Pour déterminer ce qui existe en dehors de l'expérience phénoménale, Tononi et ses collègues estime qu'il faut recourir à un critère opérationnel : l'existence de l'entité physique conduit-elle à une explication fiable et prédictive des expériences phénoménales ? À titre d'exemple, on pourrait imaginer que l'existence d'un arbre dans le monde est validée si cette idée permet de rendre compte des expériences réitérées de la table dans notre conscience (chaque fois que j'ai regardé de ce côté, j'ai fait l'expérience d'un arbre), et qu'elle permet en outre de faire des prédictions (lorsque je regarderai à nouveau de ce côté, je verrai un arbre). Dans le cadre d'une théorie systématique, les objets manipulés ne sont toutefois pas des arbres, mais des unités élémentaires : les électrons, les neutrons, les hadrons, etc. Or comme il s'agit ici d'expliquer ce que c'est que la conscience, ces unités sont vues comme des éléments irréductibles d'information neuronales.

4. **Réductionnisme opérationnel (atomisme)** : D'après Tononi et ses collègues, il n'est en principe pas besoin de postuler autre chose que l'existence d'unités élémentaires et de relations de causalité les régissant. Du point de vue mathématique, la tâche ultime de la physique consiste dès lors à définir une matrice de transition probabiliste T_U stipulant, pour chaque état u de l'univers à un instant donné (l'état u consistant en une certaine disposition de toutes les unités élémentaires), quelle est la probabilité d'observer l'état u' à l'instant suivant. Il faut en outre que l'observateur soit en mesure de manipuler individuellement chaque unité élémentaire de façon à déterminer dans quelle mesure la modification de l'une d'entre elles impacte les autres unités du système. À ce titre, et pour conclure, on peut dire que l'épistémologie de Tononi constitue en une mise à plat des phénomènes physiques. Les concepts de loi universelle ou de propriété physique sont rendus obsolètes. N'existe plus que les entités dont la manipulation induit des différences observables dans la prédiction de l'évolution du système. Notons toutefois que ce réductionnisme est dit *opérationnel*, et non *ontologique*, car ces unités élémentaires ne sont pas nécessairement vues comme des réalités *du monde*, mais plutôt comme des composantes nécessaires à la théorisation. Tononi défend en effet que seule existe en elle-même la conscience et les processus d'informations lui correspondant, et non pas les neurones individuels (qui existent pour la conscience, c'est-à-dire de façon extrinsèque).

5. **Perspective intrinsèque** : Pour rendre compte de l'expérience consciente en des termes physicalistes, l'existence d'une chose doit être évaluée depuis la perspective intime du sujet, et non pas en prenant le

point de vue d'un observateur extérieur. Dans les termes que j'ai introduit, cela signifie qu'il faut prendre en compte une perspective en Je et non en Il. Cette dernière est conçue par Tononi comme un ensemble de propriétés mathématiques présentées par le système physique. Cela implique notamment que l'information disponible par l'esprit doit dépendre de son état propre, et non pas de la moyenne des états observés par un observateur extérieur. Sur le plan ontologique, Tononi va même jusqu'à dire que seuls existent les systèmes complexes présentant une densité maximale absolue d'information. Ces systèmes sont ceux correspondant aux processus conscients, et non à la gravitation des astres dans l'espace béant de la nuit ou aux dépressions climatiques agitant vainement l'atmosphère.

Bien que ces principes soient méthodiquement introduits par Tononi et ses collègues au début de leur article, ils ne s'accompagnent d'aucune justification philosophique. Ce sont des hypothèses de travail auxquelles le lecteur est prié d'adhérer, tout au moins provisoirement, avant d'entreprendre la lecture du système dans son entier. Outre l'absence de justification, il y a également présence d'un certain nombre d'angles morts. À titre d'exemple, je regrette l'absence d'une formulation du principe d'après lequel la conscience résulte d'un traitement de l'information. Cette idée semble cruciale dans la pensée de Tononi et bien qu'elle soit adoptée par un grand nombre de scientifiques, en ce compris Edelman, elle est pourtant débattue par la communauté scientifique. Parmi les partisans de cette idée, on trouve des auteurs tels que Baars^[11], Aleksander et Gamez^[12], Earl^[13], Jonkisz^[14], Fingelkurts et Fingelkurts^[15] ou, plus récemment, Giorgio Marchetti^[16].

Du reste, il pourrait être instructif de mettre en lien cette idée avec un axiome de l'expérience phénoménale : qu'est-ce qui dans le fait d'être conscient, tel que nous le vivons au quotidien, peut être mis en analogie avec un mécanisme de traitement de l'information ? Je défendrai dans le chapitre ... que la conscience vécue s'appuie sur un principe de différenciation. La

11. Baars BJ (1988), *A cognitive theory of consciousness*, Cambridge University Press, Cambridge

12. Aleksander I, Gamez D (2011) Informational theories of consciousness : a review and extension. In : Hernández C, Sanz R, Gomez J, Smith LS, Hussain A, Chella A, Aleksander I (eds) *From brains to systems : brain-inspired cognitive systems 2010*. Springer, Berlin, pp 139–147

13. Earl B (2014) The biological function of consciousness. *Front Psychol* 5 :1–18

14. Jonkisz J (2015) Consciousness : individuated information in action. *Front Psychol* 6 :1–17 et Jonkisz J (2016) Subjectivity : a case of biological individuation and an adaptive response to informational overflow. *Front Psychol* 7 :1–6

15. Fingelkurts AA, Fingelkurts AA (2017) Information flow in the brain : ordered sequences of metastable states. *Information* 8(1) :22

16. textscGiorgio Marchetti, *Consciousness : a unique way of processing information*, *Cognitive Processing* (2018) 19 :435–464

conscience n'est pas peuplée d'objets totalement stables, données une fois pour tous dans leur intégralité, mais elle procède par comparaison. Dans un jeu incessant de rapprochements, de différenciations, d'analogies et de symétries, elle découpe à même l'expérience sensible des catégories manipulables : formes, couleurs, motifs, concepts, toutes les distinctions auxquelles le langage nous a familiarisés, et que nous prenons pour des faits acquis, sont le résultat d'un mouvement incessant de la perception. J'appellerai ce principe le *principe de différenciation*. Mais dans l'immédiat, je souhaiterais pousser plus avant l'exploration de la théorie de Tononi et m'intéresser à ses axiomes de l'expérience phénoménologique.

15.2 Axiomes de l'expérience phénoménologique

Le point de départ de la démarche de Tononi consiste à définir, au moyen de cinq axiomes, les propriétés et la structure essentielle de l'expérience de conscience. Ces axiomes, à ses yeux, sont des propositions tenues pour vraies en elles-mêmes, qui ne peuvent pas être inférées d'autres propositions et constituent ainsi des points de départ pour l'inférence logique. Tononi ne donne pas de plus amples justifications à ces axiomes, mais il semble y voir le produit d'une certaine évidence intuitive. « L'existence phénoménologique est immédiate », écrit-il, « en ce sens que mon expérience est simplement présente, directement bien plus qu'indirectement : je n'ai pas besoin d'inférer cette existence à partir de quelque chose d'autre¹⁷ »

À ce titre, sa vision de l'axiomatique pourrait être apparentée à celle de Frege lorsqu'il écrit : « J'appelle axiomes des propositions qui sont vraies mais ne peuvent être démontrées parce que la connaissance que nous en avons dépend d'une source de connaissance totalement différente de la source logique, que l'on peut appeler intuition de l'espace. De la vérité des axiomes découle qu'ils ne se contredisent pas¹⁸ » L'espèce d'intuition dont il est question dans le cas de la conscience n'est toutefois pas une intuition spatiale, mais une intuition proprement existentielle.

Je vais lister les cinq axiomes en question et je les commenterai de façon à montrer que l'évidence intuitive qui est à leur origine n'est peut-être pas aussi indéniable qu'il ne peut le sembler à première vue¹⁹

- **Axiome 1 (existence)** : La conscience existe. Chaque expérience

17. TONONI ET AL., *Integrated information theory (IIT) 4.0 : Formulating the properties of phenomenal existence in physical terms*, p.3. La traduction est de moi.

18. Cf. lettre XV/3 du 27.XII.1899 in Gottlob Frege Briefwechsel mit D. Hilbert, E. Husserl, B. Russell, sowie ausgewählte Einzelbriefe Freges, hrsg. von G. Gabriel, Fr. Kam-bartel, Chr. Thiel, Hamburg, Felix Meiner, 1980, p. 9 (trad. fr. modif. in *Logique...*, p. 223).

19. La formulation des axiomes est extraite de scholarpedia et non de l'article de Tononi sus-cité, qui les expose de façon un peu trop laconique. Voir ici : ...

est réelle – en effet, que mon expérience ici et maintenant existe (elle est réelle) est le seul fait dont je puisse être sûr immédiatement et absolument. De plus, mon expérience existe à partir de sa propre perspective intrinsèque, indépendante des observateurs externes (elle est intrinsèquement réelle ou actuelle).

- **Axiome 2 (information)** : La conscience est spécifique : chaque expérience est telle qu'elle est – étant composée d'un ensemble spécifique de distinctions phénoménales spécifiques – différenciant ainsi des autres expériences possibles (différenciation). Par exemple, une expérience peut inclure des distinctions phénoménales spécifiant un grand nombre de lieux spatiaux, plusieurs concepts positifs, comme une chambre (par opposition à pas de chambre), un lit (par opposition à pas de lit), un livre (par opposition à pas de chambre). livre), une couleur bleue (par opposition à pas de bleu), des « liaisons » d'ordre supérieur de distinctions de premier ordre, comme un livre bleu (par opposition à pas de livre bleu), ainsi que de nombreux concepts négatifs, comme pas de livre bleu. oiseau (par opposition à un oiseau), pas de vélo (par opposition à un vélo), pas de buisson (par opposition à un buisson), etc. De même, une expérience d'obscurité pure et de silence est telle qu'elle est : elle a la qualité spécifique qu'elle possède (pas de chambre, pas de lit, pas de livre, pas de bleu, ni aucun autre objet, couleur, son, pensée, etc.)). Et étant ainsi différent d'un grand nombre d'expériences alternatives que j'aurais pu vivre mais que je ne vis pas réellement.
- **Axiome 3 (intégration)** : La conscience est unifiée : chaque expérience est irréductible à des sous-ensembles non interdépendants et disjoints de distinctions phénoménales. Ainsi, je fais l'expérience d'une scène visuelle entière, et non du côté gauche du champ visuel indépendant du côté droit (et vice versa). Par exemple, l'expérience de voir le mot « PARCEQUE » écrit au milieu d'une page blanche est irréductible à une expérience de voir « ÊTRE » à gauche plus une expérience de voir « CAUSE » à droite. De même, voir un livre bleu est irréductible à voir un livre sans la couleur bleue, plus la couleur bleue sans le livre.
- **Axiome 4 (exclusion)** : La conscience est définie, en termes de contenu et de grain spatio-temporel : chaque expérience possède l'ensemble des distinctions phénoménales qu'elle possède, ni moins (un sous-ensemble) ni plus (un sur-ensemble), et elle s'écoule à la vitesse à laquelle elle s'écoule, ni plus vite ni plus lentement. Par exemple, l'expérience que je vis est de voir un corps sur un lit dans une chambre, une bibliothèque avec des livres, dont un est un livre bleu, mais je

ne vis pas une expérience avec moins de contenu – disons, dépourvue du phénomène phénoménal. distinction bleu/non bleu, ou coloré/non coloré ; ou avec plus de contenu – disons, doté de la distinction phénoménale supplémentaire hypertension/hypotension. De plus, mon expérience se déroule à une vitesse particulière – chaque expérience dure, disons, une centaine de millisecondes environ – mais je ne vis pas une expérience qui ne dure que quelques millisecondes, ni plutôt des minutes ou des heures.

- **Axiome 5 (composition)** : La conscience est structurée : chaque expérience est composée de multiples distinctions phénoménologiques élémentaires ou d'ordre supérieur. Par exemple, au cours d'une expérience, je peux distinguer un livre, une couleur bleue, un livre bleu, le côté gauche, un livre bleu à gauche, et ainsi de suite.

En guise d'exemple, Tononi et ses collègues suggèrent de se représenter un dormeur qui, soudainement tiré d'un sommeil sans rêve, fait l'expérience du mur blanc de sa chambre, de son lit à ressort et de son corps en état de semi-léthargie. Cette expérience existe immédiatement et irréfutablement, mais elle existe pour lui (axiome 1) de façon spécifique : le mur est un mur et il est blanc (axiome 2). De plus, cette expérience est unitaire, en ce sens précis que le côté gauche de la pièce n'est pas vécu, au sein de cette expérience, séparément du côté droit (axiome 3). L'expérience est également délimitée dans l'espace et le temps : le dormeur n'a pas accès à ce qui est derrière ma tête, mais je ne peux pas retrancher ce qui lui tombe sous les yeux (axiome 4). Pour finir, l'expérience phénoménologique, d'après Tononi, est structurée par un ensemble de distinctions (le mur, le lit, le corps) et de relations (le corps est sur le lit, le lit dans la chambre), mobilisant tout autant de concepts (axiome 5). Il y a dans cette dernière illustration une conception sous-jacente de la conscience qui n'est pas sans rappeler Edelman lorsqu'il disait de celle-ci : « La conscience est l'aptitude à construire une scène complexe et à réaliser des discriminations entre les composantes de cette scène²⁰ »

En dépit de cette illustration, de nombreuses questions restent en suspens. Nombre d'entre elles tiennent de la définition exacte des mots. L'interprétation de termes comme « l'existence », le « tout » ou la « structure », à teneur hautement philosophique, conduit à des incertitudes. L'axiome 3, par exemple, stipule que l'expérience doit être envisagée comme un tout irréductible, au sein duquel le vécu ne peut pas être réduit à un ensemble d'expériences séparées. Or l'axiome 5 stipule au contraire que l'expérience est « structurée » par un ensemble de « distinctions ». Opérer des distinc-

20. GERALD M. EDELMAN, *Plus vaste est le ciel : Une nouvelle théorie générale du cerveau*, Paris, Odile Jacob, 2004, p.76

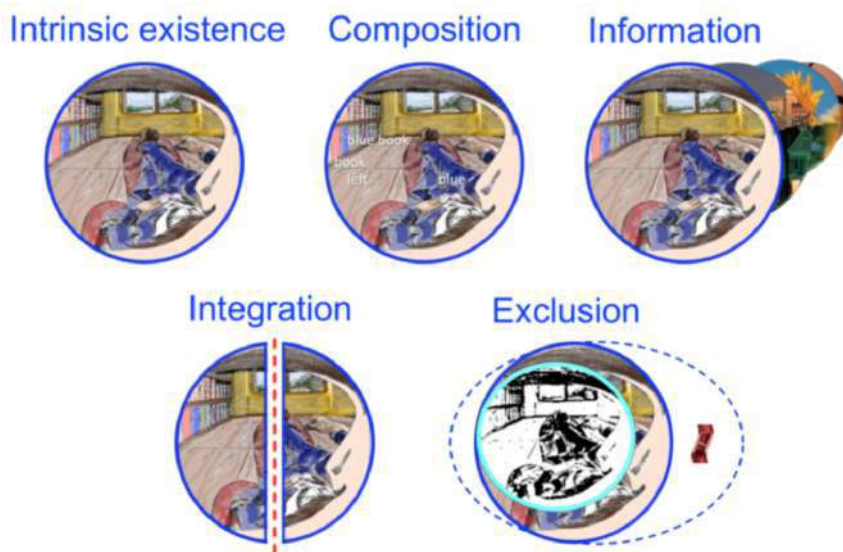


FIGURE 15.1 – Cette figure illustre les cinq axiomes de l'expérience phénoménologique à partir de la vue d'une scène.

tions, n'est-ce pas justement séparer des parties de l'expérience phénoménologique (le mur, le lit, etc...) du tout de la perception (la chambre dans son entier) ? Comment le vécu peut-il à la fois être un et multiple ? Et d'où vient exactement cette structure ?

Pour comprendre cette objection, je souhaiterais recourir à une expérience de pensée. Supposons d'un homme d'âge mûr que, suite à un traumatisme crânien spectaculaire, il ait gardé intact toutes ses facultés cognitives... à l'exception de la mémoire. Cet homme ne sait plus ce qu'il est, ni ce qu'il fait là, ni comment nommer les objets autour de lui. Si vous vous présentez face à lui, à peine vous reconnaîtra-t-il comme un de ses semblables — et aussitôt vous aurait-il oublié.

Ce damné de l'humanité, voué à un éternel oubli, est pourtant tout à fait conscient du monde qui l'entoure. Mais lorsque son attention saute d'une idée à l'autre, il oublie d'un même mouvement l'idée précédente, si bien que la vie, pour lui, se résume à un unique instant sans cesse renouvelée. Il est tantôt contemplation du paysage, tantôt il est douleur ; il est son, odeur, couleur, arbre, montagne, visages. Il s'abîme totalement dans le monde vécu, sans garder aucune notion de son existence propre. Chaque moment de la vie est naissance et mort d'un monde nouveau. C'est la création continue, sans lien de continuité entre les différents créations.

L'expérience de ce cérébrolésé est certes très éloignée de celle que nous faisons tous les jours, et peut-être même serait-il impossible de s'en faire une idée cohérente. Peut-on dire de cet homme qu'il a conscience du monde autrement que comme un tout indifférencié ? Lorsque nous inspectons, au sein

de la chambre, la présence d'une chaise, il semblerait que nous détachions cette perception isolée du tout auquel elle se rapporte. Pour ce faire, notre attention, qui se rapportait préalablement à la pièce dans son entier, doit glisser vers cette chaise singulière ; si bien que la distinction entre la pièce et les parties de celle-ci induit une temporalité.

Dans un premier temps, notre esprit est pure expérience de la pièce ; dans un second temps, il est contemplation de la chaise. La question essentielle de toute phénoménologie de la conscience consiste dès lors à se demander comment deux instants de conscience isolés se rapportent l'un à l'autre de façon cohérente, plutôt que d'être vécu comme des phénomènes déconnectés ?

Cette question n'est rien d'autre que la question de la mémoire : chaque instant de l'expérience consciente garde en lui, ou se réfère indirectement, aux instants antérieurs, tissant tout autant de lien structurels entre les idées, les perceptions et les émotions successives. C'est bien cette faculté qui nous différencie, nous autres êtres humains bien portants, du cas de ce patient cérébrolésé dont j'ai parlé plus haut.

De cela nous concluons la chose suivante : si l'expérience de conscience était réellement unitaire, elle serait telle que, au sein de celle-ci, les distinctions entre les objets, les concepts, les affects usuellement manipulés par l'esprit humain disparaîtraient. Cette myriades de catégories auxquels nous sommes familiarisés, qui découpent le réel de leurs lames aiguisées, s'effaceraient subrepticement, pour ne laisser rien d'autre qu'une appréhension indifférenciée de l'instant présent.

Sans doute est-il impossible de se représenter parfaitement en quoi consisterait une telle expérience de *pure contemplation*, pas plus que nous ne pouvons comprendre ce que serait la vie d'un être dépourvu de mémoire ? Il est toutefois envisageable que des cas limites de conscience puissent tomber sous l'observation scientifique. Je pense notamment aux déstructurations observées dans certains épisodes psychotiques. Dans son ouvrage intitulé *Dementia Praecox ou le Groupe de Schizophrenies*, Brueur défend que le symptôme primaire de la schizophrénie est l'altération des connexions liant ordinairement les associations d'idées :

“The connections between associations are lost. The disease interrupts the threads that give direction to our thoughts in an irregular fashion, sometimes affecting only a few, sometimes a large proportion of them. Thus, the result of the thought process is rendered unusual, and often logically incorrect” (p 10, translation AM).

Dans une crise de schizophrénie, la perception successive de la chambre, de la chaise au milieu de celle-ci, du mur blanc ou de son propre corps ne conduisent pas le sujet à la construction systématique d'un environnement cohérent, mais divise la conscience meurtrie en tout autant d'expérience indivisibles, où la distinction entre l'observateur et l'objet observé,

entre l'objet observé et les autres objets, émotions ou relations de la vie mentale s'efface dangereusement. Des connections existent bel et bien, mais elles sont délirantes. Elles sont, pour ainsi dire, l'ultime soubresaut d'une conscience incapable de donner une structure à son déploiement. Le sujet peut avoir l'impression d'être la chaise qu'il contemple, tout comme il peut donner à certaines perceptions des connotations affectives paranoïaques. À ce titre, on pourrait suggérer que la schizophrénie se présente essentiellement comme une maladie des concepts, en précisant néanmoins que ces concepts ne doivent pas être cantonnés au raisonnement formel, mais qu'ils concernent toute la vie mentale du sujet : la sphère logique, perceptive, affective, sociale et relationnelle. Le fait même qu'un patient paranoïaque continue, en l'absence de toute capacité à le faire, à tisser des relations cohérentes entre les choses, montre bien à quel point la conscience est, fondamentalement, une capacité à connecter et différencier des expériences entre elles.

Les axiomes 3 et 5 de la théorie de Tononi ne semblent donc pas erronés, mais ils se contentent de décrire, comme autant de faits acquis, certaines propriétés clefs de la conscience, sans s'intéresser à la dynamique liant perception, temporalité, mémoire et totalité. Ce faisant, il est impossible de comprendre, en des termes non paradoxaux, la façon dont la conscience peut se présenter, *à la fois*, comme une totalité et une fracturation. Pour le dire en court, tout laisse à penser que Tononi et ses collègues n'ont pas entrepris une réflexion de fond sur le vécu intime de la conscience. Ils semblent considérer que la vie mentale de l'individu, contrairement aux autres objets auxquels s'intéresse la science, n'a pas besoin d'être révélée au travers d'expérience spécifiques, mais qu'elle est donnée à voir pour ainsi dire directement, par le truchement magique de l'introspection.

Il s'agit là pourtant d'un statut épistémologique tout à fait unique dans l'histoire de la pensée scientifique. Pense-t-on du cerveau que l'être humain peut le comprendre en le regardant de l'extérieur ? Certes non ! On estime qu'il faut le passer sous le scalpel de la neuroanatomie, de la stimulation par électrodes, de la neuropsychologie lésionnelle ou de la modélisation informatique. Pourquoi n'en irait-il pas de même de la compréhension intime de notre conscience vécue ? Certes la conscience est un invisible, auquel personne d'autre que le sujet concerné ne semble avoir accès, mais l'exclusivité de cet accès ne veut pas dire pour autant que l'accès est immédiat. Je suggérerai dans le chapitre ... que la conscience vécue, loin de tomber sous l'inspection immédiate, requière de réaliser des expériences ; qu'elle s'appuie, pour ainsi dire, sur des concepts et des savoirs qui dépassent ce que l'on trouve en elle directement. Mais dans l'immédiat, j'aimerais m'intéresser aux autres systèmes d'axiomes existant à propos de l'expérience phénoménologie et voir s'ils contribuent, eux aussi, à cette conception un peu naïve de la conscience.

15.3 Comparaison avec d'autres systèmes axiomatiques (Aleksander & Dunmall)

Tononi n'est pas le premier à vouloir se saisir, au moyen de l'axiomatique, du soleil ardent de la conscience ; et il ne sera sans doute pas le dernier à se brûler le bout des doigts dans la tentative. Ses contemporains Igor Aleksander et Barry Dunmall, dans un article de 2003 intitulé *Axioms and Tests for the Presence of a Minimal Consciousness in Agents*, soulèvent la question de savoir s'il existe un ensemble de principes systématiques permettant de déterminer si un sujet (biologique ou artificiel) est, oui ou non, doté d'une conscience phénoménologique. De façon assez intéressante, ils proposent un ensemble de cinq axiomes assez proches de ceux de Tononi. Je souhaite examiner ses axiomes et la démarche sous-jacente, puis la mettre en comparaison avec ce qui a été dit de la théorie de l'Information Intégrée 4.0.

Dès le début de leur article, Igor Aleksander et son coauteur inscrivent leur démarche dans un cadre réaliste, supposant que la conscience est la capacité, pour un agent, à appréhender un monde sensible S qui lui fait face et à évoluer en son sein. Ce monde sensible, à son tour, est constitué d'événements, ou de changements, susceptibles d'être observés. Chaque événement peut se décomposer en un ensemble d'événements minimaux élémentaires : on peut penser à l'état de particules physiques, à la couleur des points de l'espace ou à la position des objets tombant sous les sens. Les auteurs ne précisent pas exactement ce qu'ils ont en vue, mais ils donnent l'exemple d'une télévision dont l'écran est constitué de pixel. Ils se fondent ainsi sur une ontologie réductionniste et représentationnelle, où le monde est composé d'une somme d'unités irréductibles représentable au moyen des sens. Dans ce contexte, comprendre le tout revient à comprendre précisément la relation des parties. Du reste, cela exclut la présence d'états de conscience non représentationnel, comme des idées abstraites (la justice, la vérité...), des concepts vides (les paradoxes) ou des idées dont la signification excède la représentation que l'on peut s'en faire (l'idée de l'infini).

Le monde sensible S est donc un ensemble, potentiellement infini, mais énumérable, de parties élémentaires $S_1, S_2, \dots$. Lorsqu'un changement, même infime, se produit au sein d'une de ces parties, le sujet, pour percevoir ce changement, doit lui-même accuser un changement de son état interne : l'activité neuronale se modifie, les variables d'une machine artificielle changent de valeur, etc. En d'autres termes, il existe un mapping allant de l'esprit au monde sensible. Cette correspondance n'est pas forcément linéaire ; il se peut très bien que plusieurs composantes de l'esprit changent en réaction à une seule modification du monde sensible, comme il se peut, à l'inverse, que plusieurs événements du monde soient capturés par le changement d'une seule variable interne. La relation doit cependant être injective : deux mêmes états neuronaux ne peuvent coïncider avec un même état du monde. En re-

vanche, il est possible que deux états du monde soient reflétés par le même état neuronal, mais dans ce cas, cela veut dire que ces états du monde ne sont pas distingués par le sujet. On peut songer, en guise d'illustration, au cas d'un patient daltonien incapable d'observer un changement de couleur.

Non content de se représenter le monde sensible, le sujet, d'après Igor Aleksander, possède également une capacité de se remémorer des états du monde ou de se générer ex nihilo des sensations similaires à celle que ce monde a laissée en lui. Il peut utiliser cette capacité d'imagination pour orienter son attention, prévoir le déroulement des événements, s'orienter dans le flot turbulent des événements. Ces changements internes d'états s'accompagnent d'affects, tantôt positifs, tantôt négatifs, qui l'aiguillonnent tantôt d'un côté, tantôt de l'autre, et lui servent pour ainsi dire de boussole. Ces considérations sont résumées par l'ensemble des 5 axiomes suivants :

Axiome 1 (Représentation) : Le sujet possède des états perceptuels qui représentent des parties de S.

Axiome 2 (Imagination) : Le sujet possède des états imaginatifs qui lui rappellent des parties de S ou créent des sensations similaires à S.

Axiome 3 (Attention) : Le sujet est capable de sélectionner les parties de S qu'il souhaite se représenter ou s'imaginer.

Axiome 4 (Planification) : Le sujet peut contrôler au moins partiellement la séquence des états imaginatifs de façon à planifier des actions.

Axiome 5 (Émotions) : Le sujet possède des états affectifs lui permettant d'évaluer les actions planifiées et de déterminer l'action consécutive.

Comparément à ce qui a été dit à propos de la théorie de Tononi, l'approche d'Aleksander et Dunmall a le mérite d'entreprendre une méréologie : la relation exacte entre le tout et les parties est décrite, expliquant potentiellement comment la conscience peut à la fois être le tout de la perception (S) et la perception de parties singulières au sein de ce tout ($S_1, S_2...$). Cette méréologie est cependant assez rudimentaire. Elle suppose, d'une part, que le monde sensible face part au sujet percevant (réalisme), mais aussi que la conscience soit un produit de l'activité neuronale (monisme). Pour finir, et c'est là l'hypothèse la plus questionnable, il est supposé que le cerveau, comme l'environnement, puisse être décomposés en unités indivisibles (réductionnisme).

Ces trois hypothèses qui viennent d'être exposées sont cohérentes avec les principes méthodologiques de Tononi, en particulier le réalisme (principe 2), le physicalisme opérationnel (principe 3) et le réductionnisme opérationnel (principe 4). On pourrait même être tenté de croire que Tononi opère, pour ainsi dire, une exposition plus systématique des intuitions d'Aleksander.

J'aimerais néanmoins critiquer ces principes et montrer à quels genres de problématiques ils débouchent.

Pour commencer, faisons remarquer qu'il y a une différence entre l'analogisme structurel, qui établit un lien d'identité entre la structure de l'esprit et celle des neurones, et le monisme pur et simple, qui dit de la conscience qu'elle est un produit du cerveau. C'est une chose en effet de dire que le cerveau, ou telle partie du cerveau, soit une condition nécessaire à certaines fonctions cognitives, mais cela ne signifie pas que la conscience soit pour autant « situé » des tissus neuronaux. Nous viendrait-il à l'idée que la vue soit située dans l'œil ? Certes l'œil participe au phénomène de la vision, certes la vue s'estompe lorsque les paupières sont closes, mais il est possible d'avoir deux yeux valides et de ne rien voir. La vue implique de nombreux autres organes que les sphères oculaires, allant du nerf optique au cervelet, en passant par le corps genouillé latéral, la voie ventrolatérale ou la voie dorso-latérale, si bien qu'il soit difficile de la localiser naïvement dans une région précise de l'anatomie humaine.

D'ailleurs, lorsque nous parlons de « la vue », nous supposons qu'il s'agisse là d'un concept bien défini, alors qu'il recoupe lui-même un grand nombre de distinctions. La vision des couleurs fondamentales, la détection des mouvements, la détection des formes, la vue dans l'obscurité, les saccades, etc. À chaque distinction que l'on peut opérer dans le concept de vision peut également être réalisée une distinction de fonctionnement dans les différents organes cérébraux. De là vient que le cerveau est décomposé en régions toujours plus précises, dont les interactions croissent à mesure que nous l'étudions. Mais dire que la vision se réduise à un ensemble de composantes élémentaires (couleurs, formes...) correspondant, deux à deux, aux composantes élémentaires des organes humains (cônes de la rétine, voie ventrolatérale...), est une simple hypothèse. Par seulement une hypothèse sur l'anatomie humaine ou sur la vue, mais aussi une hypothèse sur la nature langage.

Pourquoi faudrait-il que la Nature ait voulu distinguer de façon claire, dans nos organes cérébraux, la fonction de la vision du sens du toucher ? ou la vision et la perception des émotions ? Les phénomènes vécus par l'être humain sont des expériences complexes et disparates, partiellement similaires et partiellement distinctes les unes des autres, auxquelles les concepts que nous utilisons tentent d'apporter une cohésion. Or il est difficile d'isoler un pur acte de vision d'un pur acte de sensation. Tout cela se mêle en nous si intimement que, sans les concepts, nous n'aurions peut-être pas pu distinguer la vue d'un fruit succulent de la faim que celui-ci suscite — et aussi n'aurions-nous pas cherché des zones distinctes, dans le cerveau, pour la vue d'un objet en général et pour le traitement des émotions en général.

Les concepts de notre langage dressent des frontières qualitatives entre des expériences qui pourraient reposer, ultimement, sur des mécanismes anatomiques voisins. La carte de lecture des phénomènes du cerveau, telle qu'elle

serait écrite par la nature, serait-elle intelligible dans le langage de l'être humain ? Et un être doté d'une intelligence supérieure n'aurait-il pas pu se passer des concepts eux-mêmes ?

Je montrerai plus loin dans cette thèse que les concepts que nous manipulons, en ce compris la « vision » et le « cerveau », nécessitent pour être rendus intelligibles d'être réduits en unités élémentaires (les neurones, les points de l'espace, les couleurs...) qui se comprennent les uns dans leurs rapports aux autres ; si bien que l'ensemble du savoir humain est un vaste jeu de miroirs constitués de comparaisons, de rapprochements, de distinctions et d'analogies, qui nous donne la conviction qu'il existe un lien direct entre les mots utilisés et les phénomènes en œuvre dans la nature. J'argumenterai cependant en faveur d'une vision holistique, au sein de laquelle ces unités ne sont finalement que des abstractions, certes nécessaires, mais réductrices, des phénomènes envisagés dans leur interdépendance. Il y a donc une alternative au réductionnisme opérationnel défendu par Tononi, Aleksander et Dunmall.

Ce rapport complexe existant entre la structure du langage, la structure du cerveau et la structure du monde observable concourt peut-être au fait que les neurosciences, prenant au sérieux les résultats empiriques accumulés, ont dû abandonner la conception purement localisationniste, qui consistait à appairer chaque fonction cognitive (la vue, le langage, la mémoire...) à une partie du cerveau (le cervelet, l'aire de Broca, l'hypothalamus...) ? Depuis les études menées par Pierre Jean-Marie Flourens sur la façon dont des pigeons, après une lésion dans leur cerveau, restructuraient le fonctionnement de leurs aires cérébrales, ou les travaux de Hughlings Jackson sur l'épilepsie, on reconnaît l'existence d'une grande plasticité neuronale. Les cellules nerveuses ont le pouvoir de créer des associations entre des fonctions cognitives, des fonctions motrices et certains types de stimuli.

La plasticité neuronale rend caduque la volonté de dresser une carte bi-univoque associant chaque composante du phénomène de la « vision » à un organe humain, mais la question peut être posée de savoir si l'esprit est situé dans l'ensemble du cerveau ? dans des structures précises du cerveau ? dans l'ensemble du corps, entendu comme une unité psychophysique ? dans l'interaction de l'individu avec l'environnement ? dans la Nature tout entière à la façon du panvitalisme du romantisme allemand ? ou peut-être que la notion même de lieu, appliquée au concept de conscience, perd insensiblement son sens ? Cette question n'est autre que celle de la matérialité de l'esprit. S'il est quelque chose de matériel, où se trouve-t-il ? et pourquoi mon esprit n'a-t-il pas accès aux phénomènes se déroulant, présupposément, dans l'esprit des autres individus ? ou aux phénomènes ayant lieu ailleurs dans l'environnement ? Et si l'esprit n'est pas matériel, quelle est sa nature ?

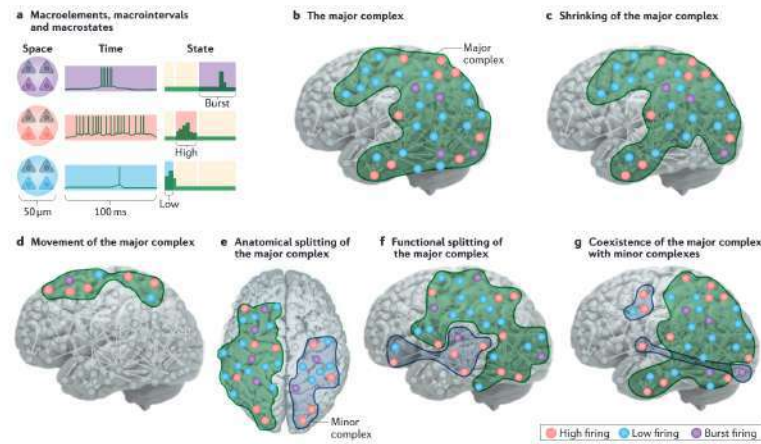


FIGURE 15.2 – Source : TONONI G. et al., *Integrated information theory : From consciousness to its physical substrate*, dans *Nature Reviews Neuroscience*, Mai 2016 p. 7.

15.4 Postulats physicalistes

Tononi affirme que la conscience, loin de reposer dans un ensemble de neurones invariables, est un processus de traitement de l'information. Il introduit l'hypothèse d'un cœur dynamique, au sein des processus cérébraux, qui coïncide avec l'état conscient. Ce cœur comprend généralement le système thalamocortical, mais il est susceptible de s'étendre, de se rapetisser, de se mouvoir dans le cerveau, voire même de se scinder ou de disparaître. Il peut donc arriver que des neurones périphériques participent de ce processus d'information fortement intégré. « Nous prédisons, écrit Tononi, que la complexité neuronale [du cœur dynamique] devrait être beaucoup plus élevée pendant l'éveil et le sommeil paradoxal que pendant les étapes profondes du sommeil lent, et qu'il devrait être extrêmement faible pendant l'épilepsie malgré l'augmentation globale de l'activité du cerveau²¹. »

La question revient dès lors à déterminer les critères permettant d'isoler ce cœur dynamique, appelé aussi le *substrat physique de la conscience* (PSC). Idéalement, il doit s'agir d'une approche susceptible d'une corroboration ou d'une réfutation empirique. Pour y parvenir, la stratégie générale consiste à prendre chacun des cinq axiomes de l'expérience phénoménologique et les associer, chacun d'entre eux, à un critère mathématique, appelé *postulat physicaliste*. Les neurones sont modélisés par un ensemble de nœuds au sein d'un graphe, susceptibles de se trouver, à chaque instant, dans deux états possibles : « activé » ou « désactivé ». Deux nœuds sont liés lorsqu'il existe un effet causal entre les neurones qu'il représente ; ou autrement dit, lorsque

21. *Consciousness and Complexity*, p. 1850

l'état d'un neurone influence statistiquement l'état d'un autre neurone.

Du point de vue mathématique, la probabilité qu'un certain neurone u s'active dépend de la somme pondérée des activations des neurones incidents $u_1, u_2, \dots, u_n$ d'après une fonction logistique :

$$\frac{1}{1 + e^{-k \sum_{i=1}^n w_i u_i}}$$

où

- Chaque neurone incident u_i est soit dans l'état activé ($u_i = 1$) soit dans l'état désactivé ($u_i = 0$).
- Le poids $0 \leq w_i \leq 1$ représente la force de la liaison liant le neurone u_i au neurone u .
- La somme des poids $\sum_{i=1}^n w_i = 1$.
- Le paramètre k représente la pente de la fonction logistique. Dans les exemples envisagés par Tononi, il est souvent égal à 4.

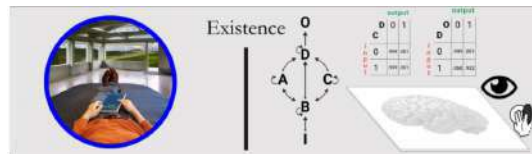
Cette fonction capture simplement le fait que l'activation d'un neurone isolé est donnée par une somme de processus de Poisson, c'est-à-dire de variables binaires pouvant soit être dans un certain état avec une probabilité donnée (neurone actif), soit être dans un autre état (neurone inactif). Un neurone va donc s'activer selon une probabilité qui est proche de 1 si tous ses voisins sont actifs, et qui est égale à 1/2 lorsqu'aucun d'entre eux ne l'est.

À partir de cette équation, il est possible de dériver la matrice de transition d'états $\mathcal{T}_U$, qui décrit à chaque instant la probabilité que le système, à partir d'un certain état s donné, évolue vers un état u^* quelconque. Par « état » du système, on entend l'assignation, pour chacun des neurones, d'une valeur égale à 1 (si le neurone est actif) ou 0 (s'il ne l'est pas). La matrice se présente comme une table qui assigne, à chacune état donné, la probabilité d'évoluer vers un autre état.

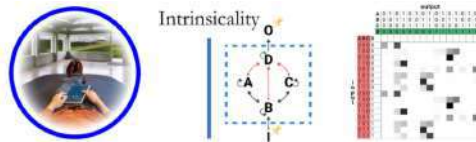
La matrice de probabilité permet de construire un certain nombre d'outils quantitatifs cherchant à mesurer tantôt la quantité d'information brute (entropie), tantôt l'information acquise détenue par une partie du système lorsque l'on contrôle la valeur d'autres parties du système (entropie conditionnelle), tantôt la distance statistique séparant deux neurones (entropie relative), ou encore, l'indépendance statistique d'un groupe de neurones par rapports à d'autres (information mutuelle). En manipulant ces outils à bon escient, Tononi va s'évertuer à construire, pour chaque axiome phénoménologique de la conscience, un critère empirique. Je me contenterai ici de donner les grandes idées théoriques : ²².

22. Ces postulats sont formulés dans tous les articles de Tononi, avec des variations en termes de précisions, de technicité ou d'argumentation philosophique. Je reprends ici

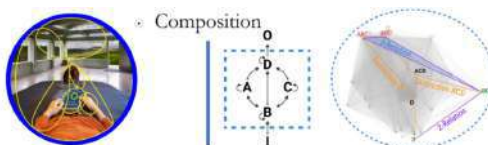
- **Existence** : Mon expérience existe, immédiatement et incontestablement. De cette régularité, j'infère l'existence d'un monde indépendant de moi. J'appelle cela « physique » si je peux observer (l'œil) et manipuler (la main) de manière opérationnelle un substrat (par exemple un cerveau) dont les unités peuvent « induire une différence et subir une différence ». Cela donne des probabilités conditionnelles qui capturent son pouvoir de cause à effet, résumé par un modèle de substrat.



- **Intrinsicabilité** : Tout comme mon expérience existe pour elle-même, le substrat physique de la conscience (PSC) doit avoir un pouvoir de cause à effet pour lui-même. Par exemple, dans ABCD, D est affecté par A, B et C, et s'affecte lui-même. La ligne bleue en pointillés indique un PSC candidat dont les pouvoirs causaux peuvent être analysés comme une matrice de probabilité de transition (TPM), avec des conditions de fond (I, O) fixes.



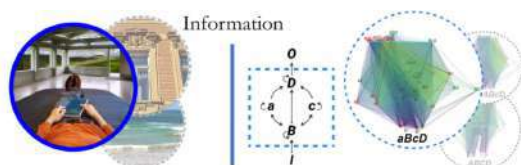
- **Composition** : Tout comme mon expérience est structurée, le pouvoir cause-effet du PSC est composé de distinctions et de relations causales qui forment une structure cause-effet (estompée). Les distinctions (orange) comprennent une cause (rouge) et un effet (vert) liés par un mécanisme (noir). Les distinctions sont liées par des relations (violet, bleu) si leurs causes et/ou effets se chevauchent.



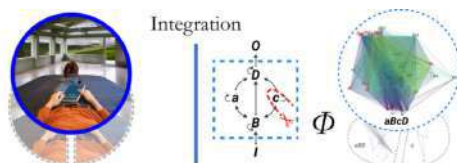
- **Information** : Tout comme mon expérience est spécifique, le pouvoir de cause à effet déployé à partir d'un PSC dans un état spécifique

la version de l'article *Only what exists can cause : An intrinsic view of free will*, qui me semble un bon compromis entre clarté, brièveté et formalisme.

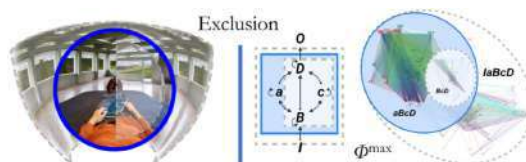
(aBcD – majuscule ON, minuscule OFF) est une structure cause à effet avec des états de cause à effet spécifiques. Et tout comme mon expérience, étant spécifique, peut être différenciée d'innombrables autres expériences, la structure cause-effet spécifiée par d'autres états de la CPS (par exemple, ABcD et ABCD) peut l'être également.



- **Intégration** : Tout comme mon expérience est unitaire, le pouvoir causal du PSC doit être unitaire. Si tel est le cas, on ne peut pas partitionner ce substrat sans perdre de l'information intrinsèque. Pour le dire en des termes simples, la somme des informations contenues dans les parties n'est pas égale à l'information du tout.



- **Exclusion** : Tout comme mon expérience est certaine, le pouvoir de cause à effet de la CPS doit être certain. Selon le principe d'existence maximale (voir texte), la frontière du PSC est l'ensemble des unités, sur un substrat donné, qui existe le plus (ici, aBcD, qui a un Φ_{maximal}), à l'exclusion de tous les autres ensembles qui se chevauchent (par ex. , BcD et IaBcD).



15.5 Comparaison avec l'information de Shannon

Bien que la notion d'information joue un rôle central dans la Théorie intégrée de l'information, elle diffère de manière distincte de celle développée dans la théorie de la communication de Shannon. Le passage qui suit cherche à faire une brève distinction entre les deux paradigmes du point de vue épistémologique. Il est supposé que le lecteur ait une connaissance au moins

générale de la théorie de Shannon. Dans le cas contraire, je le prierai de passer cette section ou de se référer à l'annexe [16.2.2](#).

Dans la théorie de l'information de Shannon, l'information quantifie la précision avec laquelle les signaux d'entrée peuvent être décodés en connaissant les signaux de sortie transmis à travers un canal (bruit). L'information (mutuelle) représente la dépendance statistique moyenne entre deux ensembles de variables, tels que les entrées et les sorties d'un canal. En général, cela dépend de l'entropie des entrées et de la capacité des entrées à prédire les sorties. L'information (auto)-contenue dans un état unique d'un système est simplement le logarithme négatif de sa probabilité d'occurrence. Comme l'a clairement reconnu Shannon, l'information dans la théorie de l'information est dissociée de la signification : l'émetteur et le récepteur sont supposés savoir ce que signifient les messages, et la théorie de l'information se préoccupe uniquement de la fiabilité et de l'efficacité de leur transmission à travers un canal. Bien que l'information de Shannon puisse être évaluée entre les états passés et présents du même système, elle est toujours évaluée du point de vue extrinsèque d'un observateur qui évalue la dépendance statistique entre les entrées et les sorties. Par conséquent, on pourrait qualifier l'information de Shannon d'information extrinsèque.

En revanche, dans l'IIT, l'information est intrinsèque : elle est évaluée du point de vue intrinsèque d'un système en termes de différences qui ont une importance pour lui. L'information intrinsèque est causale et doit être évaluée en perturbant un ensemble d'éléments de toutes les manières possibles, pas simplement en les observant. Elle est également compositionnelle, dans la mesure où différentes combinaisons d'éléments au sein d'un système peuvent spécifier simultanément différentes contraintes. De manière cruciale, dans l'IIT, l'information doit également être intégrée : si la partition d'un système ne fait aucune différence pour lui, il n'y a pas de système à proprement parler. De plus, l'IIT ne considère que l'information qui est maximale et exclusive, car, du point de vue intrinsèque, il ne peut y avoir qu'un seul ensemble de causes. Aucune de ces exigences n'a d'importance du point de vue extrinsèque d'un observateur qui mesure la dépendance statistique globale entre les entrées et les sorties.

Il convient également de noter qu'à l'intérieur du cadre de l'information de Shannon, la différence entre deux distributions de probabilités est évaluée en fonction de la divergence de Kullback-Leibler (KLD), qui mesure la perte d'information de Shannon du point de vue d'un observateur externe si une distribution de probabilités est approximée par une autre. Dans le cas où une distribution est évaluée par rapport à une entropie maximale (tous les états également probables), la KLD mesure la réduction de l'incertitude. Comme l'information dans l'IIT vise à capturer des différences qui ont de l'importance du point de vue intrinsèque d'un système, elle utilise une véritable métrique (EMD) sensible à la distance relative entre les états du système.

Plus important encore, la théorie classique de l'information ne traite pas de la signification, mais seulement de la manière dont les messages sont communiqués et stockés. Dans l'IIT, en revanche, l'information est la signification : plus précisément, l'information intégrée intrinsèque est une structure d'information maximale irréductible (MICS). Cette structure d'information intégrée capture toutes les contraintes sur le passé et le futur d'un complexe telles que déterminées par l'état de ses mécanismes. Il s'agit donc d'une forme dans l'espace conceptuel, et non pas d'un message transmis à travers un canal. En conséquence, la signification réside dans la MICS elle-même, et la signification de chaque concept individuel au sein de la MICS est auto-générée, autoréférentielle et holistique : elle est construite par les éléments du complexe, sur les éléments du complexe, et dans le contexte fourni par les autres concepts au sein de la même MICS.

À titre d'exemple, considérons la décharge d'un neurone facial profondément situé dans le cerveau et faisant partie du complexe principal. Du point de vue extrinsèque d'un observateur, un tel neurone facial peut certainement transmettre une information extrinsèque de Shannon sur certains événements dans l'environnement, étant donné que sa décharge est fortement corrélée à la présence de visages. Du point de vue intrinsèque, cependant, la décharge de ce neurone est significative en raison de la manière dont elle modifie la forme de la MICS générée par le complexe principal (qu'elle modifierait différemment si elle était silencieuse). En corollaire, les rêves ont autant de signification que les expériences éveillées : la décharge d'un neurone facial dans un rêve spécifie la même information intrinsèque même si elle ne transmet aucune information extrinsèque sur l'environnement.

En fin de compte, bien sûr, les circuits qui génèrent une signification proviennent, se développent et se perfectionnent au cours d'un long processus d'évolution, de développement neuronal et d'apprentissage, sous la pression sélective d'un environnement complexe. Selon l'IIT, l'objectif de ces circuits n'est pas tant de traiter l'information extrinsèque de l'environnement. Au lieu de cela, il s'agit de générer des structures d'information intégrée qui "correspondent" aux régularités de l'environnement de manière à ce que le "bon rêve" se produise au bon moment²³. Cette correspondance peut, à son tour, agir comme une force motrice pour l'évolution de la conscience.

15.6 Ontologie : entre éléatisme et spinozisme

Une des objections principales que l'on peut conduire à l'endroit de la théorie de Tononi consiste à se demander quelle est l'extension des choses

23. Voir notamment TONONI G., *Integrated information theory of consciousness : an updated account*, Arch Ital Biol 150 : 56-90, 2012, mais aussi TONONI G, HASHMI A, NERE A., *Sleep-Dependent Synaptic Down-Selection (II) : Single-Neuron Level Benefits for Matching, Selectivity, and Specificity*, Front Neurol 4 :148, 2013.

conscientes. À partir du moment où la conscience, en effet, est définie comme un ensemble de propriétés concernant l'information, il n'y a pas de raison, *a priori*, d'exclure de cette catégorie d'êtres des organismes non vivants comme la planète dans son ensemble, les étoiles, les processus océanographiques, voire même un simple tas de sable.

Tononi semble anticiper cette objection. Dans son essai sur le libre arbitre, publié en 2022 sur arXiv²⁴, il précise que la plupart des neuroscientifiques inscrivent leur réflexion dans une ontologie extrinsèque substantielle, tandis qu'il défend lui-même une ontologie de pouvoir intrinsèque. C'est-à-dire que, d'après la pensée dominante, ce qui existe sont les agrégats de matière, et à plus fortes raisons les unités atomiques. La réalité mentale est ainsi produite par l'articulation des neurones. Or Tononi propose à l'inverse une ontologie intrinsèque de pouvoir, où c'est bel et bien la conscience, entendue comme un système maximal d'information φ , qui est première. Toutes les autres choses, en ce compris les neurones, existent par rapport à elle de façon extrinsèque. Ils n'existent que pour elle, et non en eux-mêmes.

Du reste, il opère une distinction entre les systèmes Φ présentant un maximum relatif de ceux présentant un maximum absolu. Il précise que l'organisme humain, par exemple, présente un maximum relatif. Le système composé de l'ensemble de nos organes est plus fortement intégrés et différenciés qu'un système ne comportant que certains d'entre eux, ou d'un système incluant des objets extérieurs. Notre corps présente donc un maximum relatif. Il n'est toutefois pas comparable au degré d'intégration que l'on peut retrouver dans le cœur dynamique de l'activité neuronale, qui seule correspond à une activité consciente à part entière — c'est-à-dire, à un être intrinsèque.

« Les corps et les organes, les tables et les montagnes, les étoiles et les planètes, et bien d'autres choses qui semblent bien s'articuler sont susceptibles de se déployer en entités extrinsèques. De nombreux objets d'étude dans les sciences spécialisées, tels que les molécules, les organites, les cellules, les organes, les systèmes organiques, les organismes, les groupes, les sociétés, et d'autres semblables, qui se sont révélés être de bons moyens de sculpter la nature dans ses articulations, sont également susceptibles de correspondre aux maxima relatifs de Φ . Même si les entités extrinsèques s'articulent bien, elles n'existent pas pour elles-mêmes. Ils n'existent que par procuration, du point de vue d'une entité intrinsèque, et n'existent donc pas réellement. »

La distinction entre les maxima relatifs et absolus n'est pas exprimée de façon mathématiquement claire. L'on peut penser qu'elle réfère à la différence qui est faite, en analyse, entre un maximum local et un maximum global. Le premier est un point x tel que l'image de ce point est supérieur

24. TONONI G. et al., *Only what exists can cause : An intrinsic view on free will*, arXiv, Juin 2022

à l'image de tous les autres points dans le voisinage. Le second est un point dont l'image est supérieur à tous les autres points du domaine.

Je pense qu'il serait toutefois plus juste d'adopter, sur les propos de Tononi, une lecture spinoziste. Existe maximale ce qui est susceptible d'avoir le plus d'effets causaux sur les autres systèmes d'informations et d'être en retour impacté par eux. La théorie intégrée de l'information reprend d'ailleurs à son compte le principe éléatique, tel que formulé dans le *Sophiste* de Platon^[25] : « Je dis que tout ce qui possède une sorte de pouvoir, soit de faire quoi que ce soit à quelque chose d'autre, soit d'être affecté dans la plus petite mesure par la moindre cause, même en une seule occasion, a une existence réelle : car je prétends que les entités ne sont rien d'autre que pouvoir. » Un principe similaire est exprimé par le philosophe bouddhiste Dharmakirti^[26] : « Tout ce qui a des pouvoirs causaux existe réellement. » Il faut toutefois faire remarquer que le principe éléatique est énoncé comme une disjonction (soit causer quelque chose soit être affecté par quelque chose) tandis que le principe ontologique de la théorie intégrée de l'information se présente comme une conjonction (causer et être affecté).

Le propre de l'être, en d'autres mots, c'est d'être en mesure de faire une différence dans l'ordre des choses. Est dit existant un système tel que les informations détenues sur ce système impacte de façon significative les informations détenues sur le reste du monde, et inversement, mais sans que le comportement stochastique de l'un puisse être prédit entièrement à partir de l'autre. À ce titre, un maximum d'information pourrait correspondre à une chose qui ne peut être conçue autrement que par soi, et non en vertu d'autre chose. Il est impossible de connaître la conscience d'un individu en excluant de l'observation le cœur neuronal dynamique supportant le phénomène. Inversement, les choses du monde, que l'on tient couramment pour matérielles, sont déterminées par la conscience bien plus qu'elles ne le déterminent. Une telle interprétation conduirait l'ontologie de Tononi auprès du spinozisme, et plus particulièrement des axiomes 1 et 2 de l'*Ethique*, qui disent respectivement « Tout ce qui est en soi [existence intrinsèque] ou en autre chose [existence extrinsèque] » et « Ce qui ne peut se concevoir par autre chose doit se concevoir par soi. »

Notons toutefois que les êtres dont il est question dans la théorie de Tononi ne sont pas des substances comme chez Spinoza, mais des processus. Par ailleurs, le processus d'information, qui se déploie dans l'espace physique, et le phénomène mental de conscience, sont tenus pour être une seule et même chose ; là où Spinoza, au contraire, y voit divers attributs de la sub-

25. Cooper J. Plato : Complete Works. Hackett ; 1997, cité depuis l'article de Tononi, *Integrated information theory (IIT) 4.0 : Formulating the properties of phenomenal existence in physical terms*.

26. Tillemans T, *Dharmakirti*, in Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Spring 2021 ed. Metaphysics Research Lab, Stanford University ; 2021, cité depuis le même article de Tononi.

stance une. Il ne serait donc pas exact de dire que la théorie de l'information intégrée soit une forme de spinozisme, bien que des similarités puissent être relevées.

Au nombre de ces similarités, il faut compter l'idée de Tononi d'après laquelle il ne saurait y avoir de dualisme, ni plusieurs catégories d'être, comme le mental et la matière, ou la pensée et l'espace, mais qu'il s'agit dans un cas comme dans l'autre d'une organisation de l'information. Il défend une sorte de monisme neutre, où la conscience occupe le centre de référence : « parce que j'existe réellement — en tant que grande entité intrinsèque — les neurones de ma CFP ne peuvent pas non plus exister. Ils ne peuvent pas exister en tant que constituants de mon PSC, car ce qui existe réellement n'est pas un substrat en tant que tel mais la structure Φ exprimant ses pouvoirs causaux. Et ils ne peuvent pas exister en tant que petites entités intrinsèques à part entière, car s'ils spécifient une grande entité intrinsèque, ils ne peuvent pas également spécifier un ensemble d'entités plus petites. De plus, parce que mes alternatives, raisons et décisions existent au sein de mon expérience — en tant que sous-structures au sein d'une entité intrinsèque — les substrats neuronaux des alternatives, raisons et décisions ne peuvent pas non plus exister. Si cette image est exacte, elle ne laisse aucune place à l'émergence ou à un dualisme d'aucune sorte^[27]. »

Notons, avant de finir, que lorsque le cerveau (ou tout système complexe naturel ou artificiel) est dans un état particulier, la quantité d'information intégrée Φ dépend non seulement des états activés dans le système que d'états qui ne le sont pas, mais pourraient l'être. Les neurones silencieux contribuent ainsi tout autant à l'état de conscience que les neurones actifs. Autrement dit, la conscience est un produit de l'activité intégrée dans le réseau et est mesurée par l'intégration de l'information, une propriété qui a été définie dans les structures graphiques. Savoir qu'un système n'est pas dans un certain état confère la même quantité d'information, toutes choses égales par ailleurs, que de savoir qu'il est dans cet état^[28].

15.7 Chiasmes métaphysiques et principes épistémiques

Pour conclure ce chapitre, la théorie de l'information intégrée a le mérite de montrer comment une théorie scientifique, s'exprimant dans les termes mathématiques, objectifs, d'une perspective à la troisième personne, peut

27. TONONI G. et al., *Only what exists can cause : An intrinsic view on free will*, arXiv, Juin 2022, p. 10.

28. Plus précisément, l'entropie, mesurant la quantité d'information, dépend logarithmiquement de la probabilité que l'état soit activé. Si l'état correspondant au concept « noir », par exemple, a 20% de s'activer, savoir qu'il est activé confère plus d'informations que de savoir qu'il ne l'est pas.

s'appuyer sur une théorie introspective de la conscience, qui prend au sérieux non seulement l'expérience phénoménologique, mais qui fait également de celle-ci le critère de réalité. L'intériorité devient, en quelque sorte, plus réelle que l'extériorité, bien que cette première puisse être exprimée dans les termes mathématiques de la théorie de l'information. À ce titre, la théorie de Tononi et de ses collègues est ce qui se rapproche le plus d'un monisme à double aspect ou d'un spinozisme neuroscientifique. Il y a toutefois quelques réserves à exprimer sur le plan épistémologique.

Tout d'abord, le rapport entre l'esprit et la matière est pensé sous un rapport d'identité. Tononi n'accepte pas la présence, au sein du vécu phénoménologique, d'une dimension inexplicable. Au contraire, tous les aspects de la conscience conduisent, méthodiquement, à des postulats physiques. Inversement, les postulats physiques ne sont rien d'autre que l'expression objective de ces propriétés conscientes. En dernier ressort, l'intériorité et l'extériorité sont deux niveaux d'explication équivalents. Tononi va même jusqu'à parler d'un réductionnisme opérationnel : pour théoriser le monde et identifier les processus d'information, il est nécessaire de diviser ce monde en unités élémentaires, mais ces unités n'ont pas de réalité ontologique *a priori*. Les neurones dont il est question pour parler du cerveau et de ses mécanismes, par exemple, sont des unités indispensables aux modèles de la science, mais ils n'existent pas en soi, en dehors de l'existence préalable d'une conscience. L'esprit ne se réduit pas à l'activité d'un ensemble donné de neurones, mais agit tantôt une partie du cerveau, tantôt une autre.

Cette identité stricte qui est tissée entre l'esprit et la matière dissimule un chiasme métaphysique externaliste. Il est fait l'assomption que la réalité mentale, dans toutes ses dimensions, doit avoir une expression tangible dans l'espace étendu face à nous. Or la définition de la conscience, telle qu'on la trouve chez Tononi, s'appuie sur une distinction entre un maximum absolu et maximum relatif d'information qui ne permet pas d'identifier clairement pourquoi les processus neuronaux peuvent être qualifiés de conscients et non pas le système Terre, les courants marins ou l'union de plusieurs cerveaux. Il est certes possible, en théorie, d'infirmer ou de confirmer la théorie de Tononi par voie expérimentale, mais cela demanderait, en pratique, de faire des mesures de très haute complexité. Il faudrait, ni plus ni moins, manipuler l'activité de tous les neurones afin de voir quel est le sous-ensemble neuronal le plus statistiquement indépendant. Il faudrait ensuite comparer l'information présentée par d'autres systèmes physiques tout aussi complexes. La conviction que l'on arrive, par ce biais, à la conclusion que seul le cœur neuronal dynamique présente un maximum absolu d'information s'apparente à un acte de foi.

L'on pourra toutefois objecter que la science, pour avancer, a besoin de se fonder sur des conjectures et que celle de Tononi, aussi difficile qu'elle puisse être à vérifier, n'est plus étonnante que nombre de postulats régnant allégrement dans le paradigme de la physique contemporaine. J'accorde vo-

lontiers ce point et je suis prêt à admettre que la théorie de Tononi puisse recevoir une confirmation positive. Dans ce cas cependant, cela voudrait dire qu'on détient de la conscience une représentation complète. Or toute représentation, pour la conscience vécue, n'est qu'un simple acte de conscience parmi d'autres : c'est une idée, une sensation, un contenu mental certes sophistiqués, mais qui existe *au sein* d'une conscience préalable. Le sujet qui tenterait de comprendre la conscience avec cette représentation mathématique ne saurait donc se débarrasser du sentiment, fermement ancré, que la théorie mathématique ne capture pas le fait d'être conscient à la première personne.

Ce paradoxe, comme je l'ai dit dans l'exposition des principes généraux de ma thèse (hypothèse 2), est lié à l'idée de l'infini en acte. La conscience, quelle que soit l'idée que l'on s'en fasse, semble toujours déborder la représentation. Il faut, d'après moi, admettre dans la conscience qu'il y a quelque chose de non représentable. Or comme toutes les choses de l'espace sont représentables sous des formes géométriques, cela signifie que la conscience ne peut pas être localisable, *stricto sensu*, dans l'espace. Je ne veux pas dire par là que la science ne peut rien dire à propos de la conscience. Elle est susceptible de déterminer les frontières de la conscience, c'est-à-dire un ensemble de conditions nécessaires, mais non suffisantes, à l'expression observable de la conscience : comme le fait de détenir un cerveau, d'être vivant, etc. Ces conditions ne sont cependant pas des vérités absolues, mais des corrélations vraisemblables, qui admettent que la conscience s'exprime à travers de l'espace, sans montrer qu'elle soit pour autant entièrement appréhendable en des termes spatiaux ; tout comme la présence d'un cylindre suspendu au-dessus d'une feuille de papier projette son ombre sur cette feuille, sans pour autant être contenu dans le plan de la feuille.

Un deuxième reproche que l'on pourrait adresser à la théorie de l'information intégrée vient de la vision assez statique et naïve du vécu phénoménologique. Tononi déclare, à plusieurs endroits, que la conscience est un processus, et non une chose — et c'est d'ailleurs la raison pour laquelle il ne cherche pas à la localiser dans des parties du cerveau, mais à l'identifier à un système dynamique d'informations. Or lorsqu'il décrit la conscience phénoménologique, il semble oublier quelque peu cette dimension processuelle de la conscience. On ne trouve, dans ses cinq axiomes de l'expérience vécue, aucune loi dynamique.

Contrairement à la cinétique de Newton, qui décrit les équations du mouvement, il semblerait que la conscience, chez Tononi, soit donnée toute entière dans l'instant. Il y a cette pensée que le vécu phénoménologique soit quelque chose de transparent, d'immédiat et total. Lorsque le sujet tourne le regard à l'intérieur de lui-même, il contemple sa vie mentale à la façon d'un paysage. Ce dernier peut être décomposé en tout autant de collines, d'arbres, de fleuves impétueux, que balaye un vent silencieux, mais cette anatomie naturaliste ne nous donne pas le secret du mouvement agitant les

impressions, les pensées, les sensations, tout en les connectant les unes aux autres. La conscience en effet n'est pas dans le paysage ; elle est cet œil suspendu sur tous les paysages *possibles* constituant le fil ininterrompu de la vie. Elle est cette pupille grande ouverte du poème de Victor Hugo, qui contemple à tout instant Caïn pour lui rappeler son existence de sujet piégé dans le monde :

*Ayant levé la tête, au fond des cieux funèbres,
Il vit un œil, tout grand ouvert dans les ténèbres,
Et qui le regardait dans l'ombre fixement.*

Pour étoffer la théorie de Tononi, et la rendre compatible avec un monisme à double aspect, il est donc nécessaire de s'atteler à l'élaboration d'une théorie introspective de la conscience, qui parte du vécu dans toute sa complexité et, sans simplifier ce vécu, identifie un certain nombre de principes universels. Une telle approche doit rendre compte de l'apparent paradoxe entre l'axiome 3 (intégration) et l'axiome 5 (composition) de la théorie de l'information intégrée, qui stipulent que la conscience est une unité, un tout donné d'un bloc dans la vie mentale, mais que ce tout est décomposable en parties. La théorie de la conscience que je cherche à construire doit donc être une science méréologique, étudiant l'articulation entre les différents contenus mentaux et la conscience prise dans son tout.

Comment se fait-il que la conscience, lorsqu'elle se porte tout à tour sur la rivière, le vent, les saules ployant, ne se scinde pas en conscience de la rivière, conscience du vent, conscience des saules, mais demeure bel et bien une même conscience posé sur un paysage unifiée ? Comment se fait-il, réciproquement, que le paysage doive être décomposer en tout autant d'objets, et qu'il ne soit pas compris comme un ensemble irréductible ? Il n'existe pas deux arbres pareils, ni deux rivières semblables. Pourquoi la conscience humaine ressent-elle le besoin de diviser le monde en concepts ? Voilà tout autant de questions qui restent sans réponses.

Dans la lignée de Gerald Edelman, avec qui il a publié plusieurs articles, Gulio Tononi défend l'idée que la conscience n'est pas située dans un groupe de neurones, mais dans les processus réentrant du cerveau.

Il cherche à étoffer les critères mathématiques permettant d'identifier le cœur neuronal dynamique impliqué dans l'activité conscience. Ce cœur n'est pas donné une fois pour toute : ce n'est pas, *strico sensu*, le cortex thalamocortical, ni le système limbique, le lobe préfrontal ou le réseau du mode par défaut, mais une activité neuronale susceptible de s'étendre, de rapetisser, de se déplacer, voire de disparaître momentanément, ou de se scinder.

Sur le plan méthodologique, Tononi part de l'expérience phénoménologique de la conscience, qu'il décrit au moyen de cinq axiomes. Il fait correspondre, à ces cinq axiomes, six postulats physicalistes. Il construit ensuite une mesure d'information appelée l'information intégrée.

J'ai suggéré que l'approche de Tononi, si on se permet une lecture spinoziste, se rapproche d'un monisme à double aspect, mais qu'il s'attire néanmoins deux objections :

1. Le mental est pensé comme étant identique au matériel : il n'y a rien dans la vie consciente qui ne puisse, d'après Tononi, être représenté comme des processus d'information. Il s'ensuit que le rapport entre le corps et l'esprit est un rapport d'identité stricte, et non pas un rapport de complémentarité paradoxale.
2. La vie mentale est présentée comme un paysage statique, et non pas dynamique. Pour le dire autrement, la conscience tombe sur le régime d'une intropection immédiate, transparente et totale. Il n'y a pas de mécanismes expliquant pourquoi le tout de la perception consciente se scinde sans fin en objets différenciés.

Comblar ces deux lacunes sera l'objectif de la troisième section de ma thèse, appelée « Théorie Méréologique de la Conscience. »

Chapitre 16

Fondations d'une théorie méréologique de l'esprit

S'il m'est permis de récapituler les résultats obtenus jusqu'à présent, je pense avoir suggéré que le monisme à double aspect est viable dans les théories neuroscientifiques contemporaines. D'une part en effet, il existe un courant de pensée, promu par Henri Atlan^[1] et Antonio Damasio^[2], qui s'inscrit dans une interprétation spinoziste du rapport entre le corps et l'esprit. D'autre part, j'ai montré que des auteurs comme Gerald Edelman et Giulio Tononi défendent une théorie de la conscience établissant des analogies de structure entre l'espace mental et l'espace physique. Bien qu'ils ne l'aient pas conçu comme tels, rien ne nous interdirait d'approfondir cette approche de façon à soutenir une vision bi-perspectiviste du corps et de l'esprit inspirée de l'hypothèse de Jung-Pauli.

En dépit de ces résultats prometteurs, j'ai toutefois repéré quelques angles morts dans le cadre neuroscientifique. Au rang de ceux-ci, comptons tout d'abord, chez Tononi, la présence d'une ontologie computationnaliste, d'après laquelle la conscience, et la cognition en général, sont assimilés à un processus de traitement d'information. J'ai montré qu'une telle approche s'exposait à des difficultés de principe : tout d'abord, il est difficile d'opérer une ligne de démarcation claire entre ce qui doit être conscient et ce qui ne l'est pas. Bien qu'un critère mathématique soit donné par la théorie intégrée de l'information, il est difficile de garantir que ce critère ne se révèle pas, dans la pratique, soit trivial (tout processus complexe est conscient), soit exclusif (rien n'est conscient). En l'état actuel, du reste, le critère est difficilement vérifiable, car il demanderait d'effectuer des calculs pointus sur les différentes quantités d'information intégrée présente dans les sous-composantes du cerveau.

Par ailleurs, la théorie de Tononi défend une vision de l'expérience phé-

1. ATLAN HENRI, *Les étincelles de hasard*

2. DAMASIO ANTONIO, *Spinoza avait raison*

noménologique quelque peu indigente, fondée sur la présupposition (1) que l'introspection est immédiate, (2) que la conscience est à la fois unitaire et composée de partie, (3) que la conscience peut être décrite à l'aide de propriétés statiques, sans que soit donnée la dynamique interne et (4) que la conscience se contente de traiter de l'information issue de l'environnement, sans jouer de rôle particulièrement actif dans la mise en forme du réel. C'est pourquoi j'aimerais, dans cette partie de ma thèse, jeter les fondements d'une approche méthodologique, qualifiée de *méréologique*, qui révise ces quatre dimensions.

16.1 Critique de l'empirisme cognitivisme

Si le nombre et la nature exacte des facultés constituant le royaume de la conscience fait encore l'objet de débats, il y a néanmoins un consensus implicite pour dire que les impressions³, les idées⁴, les schèmes sensorimoteurs⁵, les espaces de similarités perceptuelles⁶ et autres produits dérivés de l'esprit humain sont issus du contact avec le monde sensible. C'est du frottement avec le réel que nos outils cognitifs, au départ rudimentaires, acquièrent peu à peu le développement qu'on leur reconnaît à l'âge adulte.

L'empirisme est si prégnant dans les sciences cognitives que Susan Carrey, dans son livre *Core Cognition*, se croit subversive en disant : « [Je] remets en question une présupposition largement partagée à propos des origines ontogénétiques de la conceptualité humaine : l'idée d'après laquelle les ressources initiales de représentations mentales humaines sont limitées à des primitives perceptuelles et sensorimotrices⁷ ». Sa thèse consiste à dire que l'être humain naît avec des structures cognitives innée, qu'elle appelle, dans la lignée d'Elizabeth Spelke, la *core cognition*. Bien qu'elle ne soit pas acquise au contact de l'expérience, cette faculté n'en demeure pas moins sélectionnée en amont par l'évolution ; si bien que le consensus empiriste, d'après lequel l'esprit se sculpte au contact du réel, n'est pas ébranlé. Il est simplement transposé de l'expérience individuelle, ontogénétique, à l'expérience de l'espèce en son entier, phylogénétique.

C'est que dans la conception cognitiviste majoritaire, l'être humain, avec ou sans *core cognition*, est un animal aveugle, échoué par hasard dans notre monde hostile, et cherchant à tâtons son chemin. Les représentations auxquelles cet être accède progressivement, depuis les premiers stades senso-

3. Hume

4. Locke et Hume

5. PIAGET, J. (1980). *Language within cognition : Schemes of action and language learning*. In M. Piatelli-Palmerini (Ed.), *Language and learning*. (pp. 163–183). Cambridge, MA : Harvard University Press

6. QUINE, W. V. O. (1960). *Word and object*. Cambridge, MA : MIT Press

7. SUSAN CARREY, *The Origin of Concepts*, OXFORD UNIVERSITY PRESS, 2009, p.

27. LA TRADUCTION EST DE MOI.

rimoteurs jusqu'à l'âge de la raison, lui permettent de naviguer dans cet environnement et de s'accaparer les ressources nécessaires à la survie. Il y a certes débat sur la façon dont ces représentations sont constituées : par sélection biologique ? par apprentissage ? en suivant des schèmes ontogénétiques ? Mais quelle que soit la réponse adoptée, les concepts n'en restent pas moins façonnés par la « lutte » de l'individu, ou de son espèce, pour la survie. Les émotions elles-mêmes, au sein de ce paradigme, ont tendance à être interprétées tantôt comme des sources de motivation⁸, tantôt comme un feedback permettant à l'individu de se choisir des objectifs et d'évaluer ses acquis⁹. À bien des égards, l'être humain est conçu comme une vaste machine adaptative, acquérant des compétences utiles dans un processus continu de rétroaction du type essais-erreurs-corrections-récompenses.

Il n'est pas surprenant, en partant d'une telle épistémologie, que l'on finisse par conclure que l'esprit naisse de la matière. Puisque l'adaptation à la réalité du monde tangible a été préalablement pensé comme la condition de possibilité de toute activité mentale structurée, il n'y a plus qu'un (petit) pas à faire avant de nier l'existence autonome, et pour ainsi indépendante, de cette vie mentale. Nos contemporains sont loin de l'idéalisme de Berkley lorsqu'il écrivait : « comme il m'est impossible de voir ou sentir quelque chose sans en avoir une sensation effective, il m'est pareillement impossible de concevoir dans mes pensées une chose sensible ou un objet, distinct de la sensation ou perception que j'en ai. » L'éliminativisme qui sévit, pour le meilleur et pour le pire, dans l'univers philosophique actuel, en récusant tour à tour les notions d'états mentaux, d'intentions, de volonté et même de conscience ; cet éliminativisme est à l'origine de l'empirisme plus qu'il n'en est le résultat. Il est le principe cohérent au fondement de l'idée que « le savoir vient de l'expérience du monde ». Toute tentative consistant à réhabiliter la conscience vécue (ou conscience phénoménale) dans l'horizon scientifique, mais sans s'attaquer à l'idée pragmatique d'après laquelle les objets de la vie mentale sont façonnés par le monde matériel, est vouée aux plus grandes difficultés.

C'est pourquoi il est impératif, si l'on entend faire honneur à l'hypothèse du monisme à double aspect, de fournir une théorie de l'esprit humain jouissant de la « neutralité ontologique ». Qu'est-ce que je veux dire par là ? Que si l'on veut éviter que l'esprit ne puisse, en dernier ressort, être réduit à la matière — ou inversement que la matière soit ramenée à la perception du sujet — il faut que la théorie que je me suis donné pour tâche de construire puisse recevoir deux interprétations concomitantes. Elle doit à la fois être compatible avec la lecture empirique, qui situe la naissance des concepts dans le développement neuroanatomique ; et à la fois avec la lecture idéaliste, qui voit le monde comme une structuration interne de l'univers conscient.

8. ALAIN DAMASIO

9. Igor Aleksander et Dunmall.

Notons que j'ai utilisé le terme « idéaliste », mais j'aurais pu, plus adéquatement, utiliser le terme « transcendantal¹⁰ ». Mon approche théorique ne postule pas que le monde existe uniquement au sein de la perception du sujet ; elle déclare que l'existence ou la non existence du monde est indifférente à son propos. À la façon de Laplace's'écriant à la face de Napoléon « Je n'ai pas eu besoin, sir, de l'hypothèse de Dieu ! », mon système philosophique, s'il est bien construit, peut fonctionner sans faire de référence explicite au monde des choses en soi. Sur quoi s'appuiera-t-il dès lors ? Sur la distinction existant entre le sujet percevant et l'objet de la perception.

16.1.1 Le Problème de la genèse de la conscience

Du point de vue de la conscience vive, il n'existe pas de différence de principe entre une perception et une illusion, entre la sensation et l'imagination ou entre la vérité et la fausseté. Toutes ces manifestations conscientes en effet existent en nous ; elles sont l'expression de notre existence en tant que sujet. C'est ainsi que, pour le patient atteint du syndrome de Charles Bonnet¹¹, les hallucinations aux allures de peintures abstraites paraîtront tout aussi réelles que la vision de son propre corps — quand bien même il sait que ce sont des hallucinations. Le vécu de conscience ne fait pas de distinction *a priori* entre la perception et l'hallucination, bien que cette distinction entre le vrai, le faux, le réel, l'illusion, le tangible, le mirage, la matière sensible et les spectres de l'imagination est pourtant fermement ancrée dans notre expérience de tous les jours. La question n'est plus alors : comment se fait-il que les concepts de l'esprit humain soient adéquats au monde extérieur ? Mais bien plutôt : comment se fait-il qu'apparaisse dans l'esprit, considéré de façon solipsiste, la conviction claire que des objets de la pensée existe en dehors de la pensée qu'on en a ?

Expliquer pourquoi la conscience, en raison de son fonctionnement interne, opère une distinction nette entre l'acte de la pensée (ou le fait de concevoir quelque chose) et l'objet de la pensée (ou la chose même qui est conçue), c'est apporter une réponse à l'origine de l'intentionnalité. Cette origine, si elle s'en tient à décrire le mécanisme propre de la vie mentale, est bel et bien neutre sur le plan ontologique. Pour s'en rendre compte, supposons que l'on détienne un *mécanisme intentionnel*, expliquant la constitution, au sein de la conscience, de la distinction entre le sujet et l'objet, et montrons successivement (1) qu'un tel mécanisme n'est pas incompatible avec une vision réaliste du monde et (2) qu'il n'est pas davantage incompatible avec

10. Le terme transcendantal réfère à la théorie kantienne qui, tout en faisant l'économie de la noumène (la chose en soi), ou plus précisément en niant la possibilité d'atteindre la connaissance des choses en soi, ne réfutait pas nécessairement leur existence

11. Le syndrome de Charles Bonnet désigne la survenue d'hallucinations visuelles complexes chez des sujets âgés présentant un déficit visuel lié au vieillissement (cataracte, glaucome, dégénérescence maculaire...) mais pas de troubles mentaux.

une vision idéaliste du monde.

En ce qui concerne le point (1), il suffit de dire que la conscience a développé ce mécanisme de sorte à s'adapter aux contraintes du monde réel. Et pour cause : en termes de survie de l'espèce, il est certes plus commode — personne ne le contredira — de détenir le concept d'un monde extérieur à la pensée, plutôt que d'être replié, telle une monade leibnizienne, sur l'univers clos de la conscience. En ce qui concerne le point (2), il suffit de soutenir que c'est, au contraire, le mécanisme intentionnel de la conscience qui est responsable de la croyance, pourtant superflue, en l'existence d'un monde extérieur à nos sens.

Défendre l'interprétation (2), notons-le bien, ne veut pas dire pour autant qu'il soit impossible de faire des prédictions sur les phénomènes futurs, d'interagir d'après des schémas stables et raisonnables, ni même de se comporter respectueusement à l'égard d'autrui^[12]. Les objets mentaux, quand bien même ils ne seraient que des illusions de la conscience, n'en restent pas moins régis par des constances fondamentales. Quelle différence cela fait-il, en pratique, que les lois de la nature soient inscrites en lettres mathématiques dans le livre de la nature ou en lettres mathématiques dans le livre de la conscience ? Quelle différence concrète entre la croyance d'exister dans le monde matériel des atomes ou celle de vivre dans un rêve ? — un rêve doté d'objets, de lois, de structures et tout à fait indifférenciables du monde tenu pour réel.

Il est vrai que ces conceptions ontologiques ont un impact considérable sur des questions telles que celles-ci : quel est le sens de la vie ? qu'y a-t-il après la mort ? ou encore, qu'est-ce que le bien et le mal ? Ces interrogations, heureusement, sont de nature spéculative, tandis que le mécanisme interne de la conscience, quant à lui, relève d'un état de fait — un fait qui n'est certes pas scientifique, mais que l'on pourrait qualifier de fait existentiel, ou encore, de fait introspectif^[13]. Or il relève de la définition même des questions spéculatives que d'être indépendantes des réponses apportées aux questions de faits. Je ne nie pas ici que les faits puissent engager la croyance

12. Ces objections sont fréquemment alléguées, dans les articles neuroscientifiques, à l'encontre du réalisme. Tononi dit par exemple, de façon assez peu argumentée : « We should assume that something exists (and persists) independently of our own experience. This is a much better hypothesis than solipsism, which explains nothing and predicts nothing. » (*Op citation*, p. 6).

13. L'expression « fait introspectif » peut laisser à penser que le mécanisme de la conscience soit quelque chose d'accessible par l'inspection immédiate de notre univers mental. Or le concept de « fait » n'induit en rien « l'immédiateté ». Les vérités scientifiques, par exemple, sont bel et bien des faits, mais il n'y a rien d'immédiat au fait de créer une collision entre des hadrons invisibles pour en déduire quelque chose à propos de la gravitation. Il ya dans le dispositif expérimental quelque chose d'une autre nature que l'objet d'étude, qui lui est étranger, mais qui lui donne accès. Je défendrai de même, tout au long de cette thèse, que les faits introspectifs ou existentiels de la conscience ne tombent pas sous l'évidence, mais requièrent le recours à des concepts et à des théories qui, en elles-mêmes, n'ont rien d'introspectif.

d'un côté ou de l'autre, mais ils ne la contraignent en rien. Je dirais donc, pour conclure, que si le but de la démarche était d'affirmer l'existence d'une conscience au sein des autres espèces vivantes, et d'exhorter pour cette raison mon semblable à respecter des valeurs morales, il serait sans doute préférable de défendre l'existence du monde réel ; mais en ce qui concerne l'élucidation du mécanisme de la conscience, cette hypothèse du monde réel n'est pas nécessaire — et elle peut même lui être nuisible.

Que l'on suppose ou non l'existence du réel, la théorie de la connaissance que je me propose de développer s'intéresse donc à la conscience comme à une sorte de réceptacle. Un réceptacle traversé tantôt de pensées et de sensations, tantôt de doutes, de peurs, de plaisirs ; et contenant, en bref, la totalité *possible* de l'expérience vécue. Cette conscience est, pour ainsi dire, *le monde à part entière*. Rien ne peut être conçu, expérimenté, senti, touché, pensé, en dehors de cet espace clos. Il s'ensuit que, à un instant *t*, le sujet fait l'expérience d'être ce qu'il sent, ce qu'il voit, ce qu'il pense. Si ces sensations, ces perceptions, ces pensées ne changeaient pas au cours du temps, le sujet n'aurait aucune raison de croire qu'il existe une différence entre lui et le monde. Il se contenterait d'une immobilité parménidienne, où tout ce qui est est la conscience, et où la conscience est tout ce qui est.

Or il se trouve que les contenus de la conscience, c'est un fait expérimentable, changent au cours du temps. La conscience que j'ai du monde à un moment donné va évoluer à l'instant suivant. Au travers de cette progression, je fais l'expérience que le monde n'est pas tel que je me le représentais de prime abord, mais qu'il est toujours différent, au moins partiellement, de la connaissance que j'en ai acquise. Je croyais voir un être humain au bord de la route, je découvre avec Diderot qu'il s'agit d'une souche. Je croyais être face à un mendiant en haillon, je découvre comme Télémaque qu'il s'agit de mon père Ulysse. Cette expérience du changement, de l'erreur, de l'illusion, permet de distinguer ce que je suis moi (une capacité de conscience) de ce que sont les choses que je connais (les contenus changeants de cette conscience).

Ces considérations me conduisent à formuler quelques hypothèses à propos de la conscience : la distinction entre le sujet percevant et l'objet de la perception proprement dite ne peut être pensée en dehors du temps qui s'écoule. Ce temps, du point de vue de la conscience, c'est le changement survenant au sein des contenus mentaux. Husserl disait que la conscience est toujours « conscience de quelque chose ». Je dis quant à moi que la conscience est avant tout conscience « de quelque chose qui change ». Ce principe est ce que j'appellerai le *principe de différenciation*.

Par ailleurs, je milite pour la réhabilitation critique du pouvoir créatif et imaginaire de la conscience. La perception du monde convoque le pouvoir créatif de la conscience. Cette dernière, loin d'être seulement un réceptacle passif du monde, est un pouvoir agissant qui participe, en chaque instant, à la constitution des contenus mentaux. Une attestation de ce pouvoir agis-

sant de la conscience peut être trouvée dans nombre d'illusions perceptives, notamment l'illusion concernant la saturation des couleurs (voir figure ??).

J'aimerais penser la conscience, non plus dans une vision purement matérialiste, qui demande à l'organisme vivant de s'adapter à l'environnement, mais dans le cadre d'un monisme à double aspect supportant une double lecture : à la fois la conscience peut être vue comme un organe s'adaptant évolutionnairement aux contraires de l'environnement ; et à la fois c'est une expérience de conscience embrassant, en elle, l'entièreté du monde perçu. C'est donc à la fois un objet au sein du monde, une partie insignifiante dans le tout qui la surplombe ; et c'est à la fois le monde dans sa totalité. Il reste à se demander par quel biais cette conscience pourrait être étudiée ? Est-il possible de l'approcher comme un objet sans sombrer dans les chiasmes métaphysiques que j'ai relevé avec le paradigme neuroscientifique ? Comment s'assurer que la démarche introspective reste en dialogue avec les résultats universels de la méthodologie scientifique ?

16.1.2 Méthode d'investigation à double aspect

Si l'objectif est de tisser une analogie entre le fonctionnement de l'esprit avec la conception scientifique du monde, peut-être est-il bon d'aller chercher de l'inspiration du côté des grands principes de la physique ? La cosmologie aime à étudier le développement de l'Univers en partant du Big Bang. Elle voit l'Univers sentier dans une tête d'épingle et décrit, à partir de ce point, les trajectoires de toutes les particules. De la même façon, je voudrais trouver le Big Bang de la conscience ; l'état de fusion à partir duquel se déploie la matière dans l'espace physique et les phénomènes mentaux dans l'espace psychique. Un tel état, s'il existe, devrait être tel que la contradiction entre la matière et la pensée, l'intérieur et l'extérieur, le sujet et l'objet s'estompe. Il devrait permettre de penser l'unité existant derrière les deux perspectives.

Idéalement, cet état doit pouvoir être étudié simultanément depuis la lorgnette du scientifique et depuis le point de vue introspectif à la première personne, de telle sorte à ce que leurs perspectives convergent. Ce que j'entends par là, c'est que notre fameux Big Bang de la conscience doit surmonter les chiasmes métaphysiques. À première vue, un tel réquisit semble irréalisable — et je sens d'ici le scepticisme du lecteur posé sur ma tête comme une épée de Damoclès. Je tiens à le rassurer d'emblée : il est impossible d'isoler une pure conscience vive et une pure conscience morte. Je ne dédis pas ce que j'ai montré plus haut. Ce que je défends, en revanche, c'est qu'il est possible de tendre, de converger pour ainsi dire, vers un point où les chiasmes s'effacent. Pour clarifier à quoi peut bien ressembler ce Big Bang, adressons les deux questions suivantes : que reste-t-il du cerveau lorsqu'on réduit tendanciellement toutes ses stimulations ? que reste-t-il de la conscience vie lorsqu'elle se vide progressivement de tous ses contenus ?

Sur le plan de l'observation *scientifique*, un cerveau sans activité neuro-

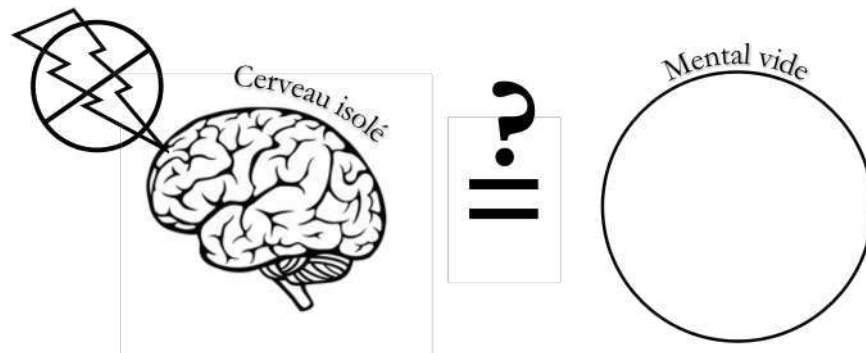


FIGURE 16.1 – Un cerveau isolé de toute stimulation est-il l'équivalent d'un mental sans contenus ?

nale — ou avec une activité très faible — serait au plus proche de la *pure* conscience morte. Elle n'aurait en effet plus rien de phénoménal en elle, plus rien de proprement vivace. Plus de pensées, de sensations, de doute. Rien que des nerfs à l'état brut. Or sans état mentaux, le paradoxe du zombie scientifique, responsable du chiasme observationnel, n'en est plus un. Je ne veux bien sûr pas me pencher sur le cas d'un cerveau cryogénisé, mais d'un individu qui, dans le cadre d'une expérience consentie, accepterait de réduire toutes les stimulations alimentant son organe cérébral. Cela pourrait être savamment réalisé en construisant un dispositif qui l'isole de toute source de lumière, d'odeurs, de bruits et de sensations tactiles.

Sur le plan de l'introspection *philosophique*, un esprit sans états mentaux — ou avec une activité tendanciellement nulle — s'approcherait sensiblement d'un *pur* réceptacle de conscience. N'est-ce pas ainsi que le Je a été défini ? Ce serait le contenant sans contenu, le bocal de l'aquarium sans eau ni aucun poisson. Pour atteindre cet état, qui s'apparente une fois encore — reconnaissons-le — à une absence de conscience, le sujet de mon expérience devrait utiliser des techniques de méditation pour apaiser son esprit. Pensées, impressions, anxiétés, tous les remous de l'âme doivent s'immobiliser dans l'ici et maintenant, pour laisser place au mystère hypothétique de la simple présence.

Si ces deux activités, la méditation par voie introspective et l'isolation des stimulations du cerveau, sont réalisées de concert, la conscience vive et la conscience morte convergent vers un même point. À l'état limite, un cerveau sans activité et une conscience sans objet sont identiques. Le Je se confond au cerveau dans sa totalité. Peut-être objectera-t-on que l'expérience dont je parle s'apparente à étudier le cerveau d'un défunt ? Que c'est de peu d'intérêt à la compréhension du phénomène même de la conscience ? J'en conviens, si toutefois l'état de mort cérébrale était seulement atteint. Or ce qui est crucial dans mon approche, c'est qu'on s'intéresse uniquement à ce qui se produit au cours du processus de convergence, sans jamais conduire

l'expérience jusqu'au bout. Comment pourrait-on seulement supprimer entièrement l'activité mentale ? Le but n'est pas de commettre un homicide mystique !

Si la conscience est bel et bien, comme je le défends, représentée par une hypersphère, il doit être rigoureusement impossible de faire coïncider la perspective en Il et la perspective en Je. De quelque façon que vous tourniez l'objet en quatre dimensions, les deux sphères projetées dans l'espace tridimensionnel ne peuvent pas se confondre. Je n'ai du reste pas cessé de dire que le Je n'était pas le cerveau dans son entier, pas plus que la conscience n'était des phénomènes dans le cerveau. L'on s'attend donc à ce qu'une activité répulsive survienne, au plus près de l'état limite, qui empêche d'opérer cette identification entre le Je et le cerveau. Cette activité répulsive doit être similaire à ce qui se produit face à la mort, quoique ce phénomène soit moins aisé à étudier.

De toutes ces considérations, je conclus qu'au moment d'approcher l'état de vide, il faut soit que des images se forment dans la conscience vive, soit que des activités apparaissent dans le cerveau. En l'absence de ce genre de manifestations, on peut raisonnablement suggérer que la conscience n'est pas correctement illustrée par l'hypersphère et que ma théorie doit être révisée. Si ma prédiction est confirmée, en revanche, il s'agit d'une corroboration. Ce n'est en aucun cas une preuve — je pense d'ailleurs qu'aucune preuve ne peut être obtenue en métaphysique. Mais cela démontre la viabilité de mon idée. Et sur le terrain dangereux de la spéculation, c'est déjà une belle victoire. La suite de mon discours se concentrera sur un cas d'étude mené par le chercheur John Cunningham — dont je décrirai la démarche dans la prochaine section. Je montrerai que cette étude de Cunningham est assez proche d'une authentique expérience mystique. Une mystique de laboratoire, pourrait-on dire.

16.2 L'Expérience du cerveau isolé

Nous sommes en 1952, dans un laboratoire de Washington. Le lieu est plongé dans un curieux clair-obscur. Dans un coin se découpent les contours d'une table jonchée de papiers écornés. On y trouve pêle-mêle des traités de neurologie, de psychanalyse et de philosophie hindoue. Par la fenêtre poussiéreuse tombe un fin rayon de lumière. Il traverse l'espace de long en large, avant de se heurter à un caisson de métal. De l'eau flotte à l'intérieur, silencieuse. Un sas permet d'y pénétrer, mais aucun son n'en sort. Le curieux sarcophage attend qu'un homme vienne s'y allonger, seul, pour le happer dans un curieux voyage. C'est muni d'un masque de respiration, spécialement conçu pour bloquer la lumière, ainsi que d'une combinaison adhérente, que des individus se succèdent dans le bassin. Sur leur visage peut se lire une étrange désinence de nervosité et de scepticisme. Mais au moment de sortir

de l'eau, le sourire est radieux. Dans cette pièce du National Institute of Mental Health, John Cunningham Lilly vient de mettre au point le premier caisson de flottaison, appelé aussi caisson d'isolation.

À l'intérieur de ce drôle d'habitable, pas plus haut qu'un homme, le sujet de l'expérience est privé de tous les points de repère, aussi bien externes que cénesthésiques, auxquels il était familier. La proprioception de son corps, en effet, est rendue très difficile au sein du liquide ambiant. Le toucher et la vue, eux-mêmes, sont sensiblement amoindris. Quant à l'ouïe, elle se limite à un lointain tambourinement. L'individu est, comme le bébé en gestation dans le liquide amniotique, perdu en flottaison dans l'espace et le temps, sans repères ni stimulations. Dans ce contexte, il lui est impossible de se raccrocher aux impressions, cognitives et perceptives, qui sont au fondement du sentiment intime d'exister dans le monde. Or que devient le cerveau lorsqu'on le soustrait à toute espèce de stimulation ?

La majorité des neurologues de l'époque, avec les psychanalystes en figure de proue, n'aurait pas hésité à répondre : le cerveau entre dans un état de repos proche du coma ou de la mort biologique. Dans son essai intitulé *Par-delà le principe de plaisir*, Freud avait en effet popularisé l'idée d'après laquelle l'organisme psychique, en l'absence de forces extérieures, cherchait à réduire le plus possible son excitation interne et externe. D'après cette conception, que sa collègue Barbara Low nomme, non sans éloquence, sous le doux nom de Principe de Nirvâna, il n'y avait pas lieu de penser qu'un cerveau isolé du bruit ambiant de l'environnement puisse conserver une activité biochimique. C'est l'aveu intellectuel que livre Freud dans son essai *Au-delà du principe de plaisir* :

« La conviction que nous avons acquise que la vie psychique, peut-être la vie nerveuse en général, est dominée par la tendance à l'abaissement, à l'invariance, à la suppression de la tension interne provoquée par les excitations (par le principe du Nirvâna, pour nous servir de l'expression de Barbara Low), cette conviction, disons-nous, constitue une des plus puissantes raisons qui nous font croire à l'existence d'instincts de mort². »

Le principe de Nirvâna découle assez naturellement de la volonté de Freud d'inscrire la psychanalyse dans une méthodologie scientifique. Et pour cause : la science, on l'a dit au premier chapitre, ne se représente pas l'esprit comme une *conscience vive*, mais comme une *conscience morte*. Le cerveau, à ses yeux, est une boîte crânienne dans laquelle des signaux sont émis en réponse aux stimuli incidents. La méthodologie scientifique consiste donc à manipuler cette boîte noire du psychisme pour en étudier les réactions, d'après le protocole rigoureusement établi — et jusqu'alors très prometteur — de la manipulation des causes et des effets.

L'idée que le cerveau puisse, en l'absence de tout input, produire une activité autonome, n'était pas facilement concevable. Mais ce qui est vérita-

blement ironique, c'est que Freud, comme s'il avait voulu donner du grain à moudre à ma définition de la *conscience morte*, parle justement d'instincts de *mort* pour désigner la tendance de la machine psychique à revenir le plus vite possible à l'état inertiel (ou inerte). « L'être vivant élémentaire », écrit-il un peu plus loin dans son essai, « serait très volontiers resté immuable dès le début de son existence, il n'aurait pas mieux demandé que de mener un genre de vie uniforme, dans des conditions invariables. »

D'après Freud et ses disciples, le psychisme est un organe immobile qui ne cède en rien à l'Être parméniidien³. C'est un monde clos, pétri du désir de revenir à son état originel d'être unicellulaire, qui n'aurait guère évolué sans le concours d'un jeu de forces étrangères. Il ne tarderait d'ailleurs pas à mourir si l'environnement extérieur, et les lois retorses de l'évolution, ne le forçaient à prendre des détours sur le long chemin menant de la naissance à l'extinction. Loin d'un organisme qui, mû de sa propre force, s'accroît en dépit des contraintes mortelles qui s'exercent sur lui, le cerveau de la psychanalyse est un organe qui cherche, dans le labyrinthe de la vie et des pulsions sexuelles, un raccourci vers le repos de la mort.

C'est dans ce climat quelque peu mortifère, dominé par un esprit de corps tenace, que John Cunningham, grâce à son caisson d'isolation, parvient non seulement à donner une dimension empirique à la question de l'activité du cerveau en l'absence d'excitations, mais aussi à démontrer à la communauté scientifique, tout d'abord sceptique, que le principe de Nirvâna doit être révisé. Les sujets sortants de son laboratoire décrivent certes un sentiment de paix intérieure, mais la période initiale de repos dans le caisson de flottaison est peu à peu remplacée par une activité hallucinatoire digne des grandes extases du mysticisme chrétien. Tout se passe, en dernier ressort, comme si le cerveau, abandonné à cet état d'homéostasie radicale, se mettait en devoir de reproduire les stimulations auxquelles le dispositif expérimental l'avait savamment soustrait.

C'est ainsi que des perceptions imaginaires, des pensées oniriques, des désirs fantasmatiques, viennent prendre la place de la perception usuelle du monde. Il en va du caisson d'isolation un peu comme du syndrome de Charles Bonnet ou du syndrome de l'organe musical : lorsque les sens de la vue ou de l'audition sont amoindris, il peut arriver chez certains sujets, ne présentant pas de lésions cérébrales ni de troubles mentaux, que des hallucinations perceptives ou auditives complexes se forment dans leur esprit. Charles Bonnet, qui fut le premier à décrire ce trouble, rapporte que son grand-père, percevait des voitures attelées, des tapisseries, des personnages, des oiseaux, des motifs en échafaudages dignes des jardins suspendus de Babylone⁴.

Que serait un cerveau sans activité ? Sans doute l'analogue d'un paysage sans ciel et sans lumière, un espace béant, angoissant, sans aucun souffle vital pour ensemençer les horizons virginaux. Le sujet, pour exister en tant que *conscience vive*, et non seulement comme un vivant appelé par la mort,

éprouve la nécessité de peupler son univers mental. L'immobilisme de la perception, l'arrêt sur image, serait pour lui synonyme de la cessation de l'existence psychique — et le renverrait peut-être à un état proche de la mort ? Il n'y a en tout cas pas de différences notables, en ce qui concerne l'intensité de l'activité psychique, entre l'état usuel de l'esprit en interaction avec son l'environnement et l'état hallucinatoire d'un esprit coupé du monde.

Quelle révolution copernicienne que cette reviviscence des cellules neuronales ! Dans le paradigme traditionnel de la biologie, la propriété essentielle du cerveau, sélectionné par des années d'évolution, consistait à identifier, à anticiper et à pallier aux menaces de l'environnement. Dans la nouvelle approche, l'idée que le rôle de l'esprit soit, non de survivre dans le monde, mais de survivre à l'absence du monde, est un retournement de perspective d'autant plus renversant qu'il déplace le centre d'attention : le vivant n'utilise plus uniquement la conscience pour appréhender le monde ; mais la conscience est un monde en soi que le vivant cherche à alimenter. On passe ainsi d'une ontologie majoritairement réaliste, au sein de laquelle le monde est réel et l'esprit est la perception de celui-ci ; à une ontologie plus idéaliste, où c'est l'esprit qui est réel et où le monde n'en est que le produit.

On pourrait bien sûr rétorquer, et cela se fait régulièrement, que les hallucinations du cerveau ont une fonction évolutionnaire. Que veut-on dire par là ? Que l'esprit, lorsqu'il n'est pas soumis aux contraintes de l'environnement, présente une activité spontanée lui permettant d'affiner ses modèles de compréhension du réel. Le sommeil, le rêve, l'imagination, d'après cette vision biologique du monde, seraient ainsi des outils mis à disposition de l'individu pour simuler des dangers et se confronter à certains contextes émotionnels.

Il en irait dès lors du cerveau comme d'une machine désœuvrée. Dans notre monde capitaliste, il n'est généralement pas bien reçu de rester les bras ballants. Si le cerveau n'a rien à faire, ne doit-il pas trouver, par lui-même, le moyen d'exercer ses talents ? Or que fait-il de mieux, sinon s'entraîner à affronter le réel ? L'usage de termes renvoyant, peu ou prou, au vocable informatique, abonde généralement dans cette espèce d'explication tout à fait instrumentale des rêves, qui assimile indirectement le cerveau à une machine d'entraînement dans un univers de réalité virtuelle. « Dans de nombreux cercles scientifiques, la croyance selon laquelle le cerveau est un ordinateur reste largement répandue. » disait Gerald Edelman après avoir gagné le prix Nobel en médecine⁵.

L'on aurait toutefois tort de disqualifier cette théorie du cerveau-simulateur sur un simple malentendu terminologique. S'il n'est pas sans danger de prendre pour modèle les outils techniques que nous utilisons (l'ordinateur) pour décrire l'objet de l'expérience (le cerveau) la naïveté n'est toutefois pas un péché en science — et vaut souvent mieux qu'une sophistication digne des scolastiques ! Bien que l'on puisse douter que le cerveau soit un ordinateur, il faut reconnaître un certain crédit à la conception biologique du cerveau.

Nous faisons régulièrement l'expérience d'un apprentissage inconscient dans l'état de sommeil. Cela se produit lorsque nous cristallisons des connaissances répétées peu de temps avant de nous endormir ou lorsque, face à un problème tenu pour impossible, nous viennent dès le réveil des hypothèses nouvelles. Certains sportifs de haut niveau, du reste, utilisent le caisson de flottaison avant les rencontres importantes. Cela les aide à améliorer leur concentration, leur présence, leur confiance.

Il est donc tout à fait possible, et même assez probable, que l'explication évolutionnaire du cerveau soit exacte, mais cela ne fait pas d'elle une explication vraie. La vérité, en effet, ne se contente pas de fournir une raison suffisante ; elle demande encore de chercher les réponses alternatives. Or je pense que la question essentielle soulevée par l'expérience de John Cunningham ne consiste pas *uniquement* à se demander si le cerveau produit ou non une activité en l'absence de stimulation — et si cela répond d'une nécessité évolutionniste. Il s'agit aussi de saisir l'opportunité de s'intéresser, parallèlement à la question sur le cerveau, à la nature du *Je* de la conscience.

Pour comprendre ce versant du problème, qui n'est pas un versant proprement scientifique, mais philosophique, il faut se mettre du point de vue de la conscience vive. À ses yeux, le monde n'est rien d'autre qu'un phénomène de conscience dont la réalité en dehors de nos sens peut être questionnée. Pour être plus exact, il est indifférent que ce monde existe ou non au-delà de la perception qu'on en a, puisque l'esprit n'aura jamais accès à rien d'autre qu'à ces perceptions proprement dites. Une question du type : « Le cerveau produit-il une activité en dehors de stimulations extérieures ? » ne saurait dès lors guère avoir de sens. La vraie question est : « Est-il possible d'être une conscience sans contenu mental ? »

Dans l'expérience usuelle de conscience, celle que nous éprouvons en dehors de tout caisson d'isolation, nous sommes un esprit peuplé d'idées, de sensations, d'émotions, d'impressions tantôt vives, tantôt imperceptibles, mêlant simultanément les sens de la vue, de l'ouïe, du goût, du toucher et de l'odorat. Or au moment de nous couper de tout rapport avec le monde que nous connaissions, et d'entrer dans l'isolation, un certain nombre de contenus sont altérés. Nos perceptions, nos idées, nos sensations ne sont plus ce qu'elles étaient. On pourrait même craindre qu'elles ne disparaissent tout à fait.

N'était-il pas permis aux contemporains de John Cunningham d'envisager, tout au moins hypothétiquement, que l'isolement ait pour effet de faire taire le mental ? Il arrive à certains sujets, peu aguerris, de s'endormir dans un caisson de flottaison. Encore que le sommeil soit usuellement peuplé de songes, rien ne nous interdit de penser, avant d'avoir réalisé l'expérience, que le caisson d'isolation ait sur la conscience l'effet d'un sommeil sans rêve — ou d'un coma comparable aux salles d'opération. Le cas échéant, cela signifierait que la conscience serait mise en suspension. Ce serait comme une petite mort de l'esprit, analogue, peut-être, à ce à quoi nous destine la fin

de notre trajectoire consciente ?

Cette question, notons-le bien, ne revient pas à se demander s'il y a une activité dans le cerveau. Pour le Je de la conscience, en effet, le cerveau objectif n'existe pas ; et s'il existe, il est inobservable. Tout au plus peut-on avoir accès à l'idée du cerveau, mais non à la chose en soi. Si la préoccupation philosophique que je décris maintient donc bien une analogie évidente avec le versant scientifique de la question, la formulation ne saurait lui être substituée. Cette formulation consiste à se demander s'il est possible d'être une conscience sans rien avoir à penser, à sentir, à éprouver ? Et si jamais cela devait être exclu — comme c'est en effet le cas — par l'expérience du caisson d'isolation, il reste à expliquer pourquoi notre mental déploie un tel effort, contrairement au principe du Nirvâna, à éviter l'extinction des sensations ?

En vue de répondre à cette grave interrogation, il s'agira d'abord de se demander à quoi ressemblerait une conscience vive qui n'aurait plus aucun sujet de sensations dont se nourrir, ni interne ni externe, pas plus que des souvenirs, des rêves et des désirs à explorer. Cette situation est, pour ainsi dire, une extrapolation du caisson d'isolation — une expérience de pensée dont nous espérons qu'elle puisse nous éclairer sur la nature du mental. Elle représente, au demeurant, l'idéal que les pratiquants de l'hindouisme, du jaïnisme et du bouddhisme ont en vue lorsqu'ils aspirent au Nirvâna.

16.2.1 Je pense donc Je lie

*Le « Moi » est une simple
conscience, qui accompagne tous
les concepts.*

Heidegger

Les résultats de l'expérience de Cunningham démontrent la présence d'une répulsion entre le pôle de la conscience morte et le pôle de la conscience vive. Tout se passe comme si vider l'esprit de ses pensées, ou le cerveau de ses stimulations, revenait à marcher avec une boussole en direction du centre de l'Antarctique. La boussole, spécifiquement calibrée pour distinguer le Nord du Sud, se met à tourner sans queue ni tête lorsque nous approchons des extrémités de la Terre. L'aiguille semble être prise dans une tempête. Elle tremble, elle saute, elle patine. Le cadran explose soudainement, mais nous ne continuons pas moins d'avancer vers l'origine de ce curieux magnétisme.

Ce qu'il faut étudier, en marin téméraire, c'est l'accalmie qui suit cette tempête. J'aimerais me demander à quoi ressemblerait une conscience qui aurait été effectivement vidée de ses contenus ? Une conscience vide, je l'ai déjà dit, est irreprésentable, mais cela ne veut pas dire qu'il soit impossible de l'envisager philosophiquement. D'une part, je peux m'appuyer sur la sensation éprouvée lors des expériences de conscience limite. D'autre part, je peux tenter d'extrapoler à partir de là quelques idées abstraites. Ma démarche, je le reconnais, s'apparente à mener une expérience de pensée sur

un terrain où, justement, rien ne peut être pensé. La conscience vide est le continent rêvé par l'explorateur. La ligne d'horizon imaginée par le cartographe. Au contraire de Christophe Colomb, il y a peu de chance que nous parvenions à mettre le pied dans les Indes de l'esprit humain. En revanche, je crois qu'il est permis d'espérer qu'un cadre conceptuel puisse être donné à la dynamique interne de la conscience. J'ai montré en effet qu'un principe de vitalité orientait l'activité du mental. La psyché, d'après des lois qu'il reste à élucider, préfère halluciner que de retourner au repos. Il s'agit maintenant de comprendre si, à ce premier principe, peuvent s'en articuler d'autres ?

J'aimerais vous inviter à fermer les yeux quelques instants. Imaginez avec moi que l'expérience du caisson de flottaison se prolonge tant et si bien que nous ayons eu l'opportunité de ressasser, individuellement, chacune de nos pensées, de nos souvenirs et de nos impressions sensibles. Que cent fois soient nés des univers entiers de rêves, de fantasmagories, de doutes béants et métaphysiques, traumatiques, accompagnés de leurs cohortes mystiques de monstruosités imaginatives ; que cent fois soient morts les espaces immaculés de la pensée, des songes et de la spéculation. Que demeure-t-il alors de cet esprit qui ne trouve plus rien, ni en lui, ni en dehors de lui, à se mettre sous la dent ? Fermez les yeux et réfléchissez-y.

Vous avez votre idée ? Pour ma part, je pense qu'il ne lui restera que sa propriété la plus essentielle : à savoir, le fait d'être, en tout instant, conscience de quelque chose. J'ai appelé cette propriété le Je, mais entendons-nous bien : le Je ne désigne pas le fait d'être telle personne dans le monde, de porter tel ou tel nom, d'occuper une certaine fonction, ni même d'avoir une certaine histoire. Le Je est dépourvu d'identité, de mémoire et de narration. Il n'est rien d'autre que la constatation qu'il y a, en cet instant donné, un phénomène de conscience. Bien que le contenu de cette conscience change en permanence, le fait même d'être conscient est ce qui demeure constant tout au cours du temps.

Je suis le contour du monde

Le Je est le contour invisible du monde mental, qui supporte ce dernier sans jamais se présenter à lui. Ce n'est ni une goutte d'eau, ni un courant marin : c'est le bassin océanique. Et un poisson perdu dans cet océan, qui se demande qui il est, ne se rendra pas même compte de l'élément dans lequel il flotte pourtant tous les jours. Il devrait idéalement sortir de l'eau pour constater qu'il y a de l'eau et en comprendre la nature. Or en dehors de l'eau, il meurt aussitôt. C'est pourquoi le Je est si fondamental, mais que, en dépit de son omniprésence, nul ne peut le pointer du doigt. « Ce centre du monde, que je nomme le Je » écrivait Wittgenstein, « est le support de l'éthique⁶ ». C'est à proprement parler là l'essence de ce qu'on appelle le Mystique. « Il ne s'explique pas, il ne se dit pas, mais il se montre⁷. ».

Ce Je dont nous esquissons le portrait est présent dans plusieurs tra-

ditions philosophiques. On le trouve tout d'abord chez Descartes lorsqu'il s'exclame : *cogito ergo sum*. La démarche du père de la philosophie moderne est d'ailleurs analogue à celle que nous avons dressée ici. Encore qu'il n'ait pas inventé le caisson d'isolation (il se contentait de la contemplation immobile des taches de son plafond), il n'en a pas moins douté de la réalité de ses sens avec une telle acuité et une telle éloquence qu'il est parvenu à convaincre des générations entières de penseurs que la propriété la plus essentielle du sujet, celle sur laquelle nous ne pouvions avoir aucun doute, était le fait de penser — non pas de penser telle ou telle chose, mais d'être une substance pensante⁸.

On trouve encore la présence de ce Je dans le chapitre qui ouvre la Phénoménologie de l'Esprit de Hegel, intitulé la *Certitude sensible*. Le philosophe de Iéna envisage l'esprit naissant de l'être humain dans un rapport immédiat au savoir. Il appelle certitude sensible la capacité à concevoir des choses singulières, des *ceci* et des *cela* concrets, comme par exemple cette maison-ci, sans être en mesure de rapporter cette contemplation à un concept. Sous ce rapport, le Je qui conçoit se confond alors à l'acte de sensation. C'est d'ailleurs cette singularité qui distingue notamment, aux yeux de Hegel, la certitude sensible de la perception. « Cette certitude sensible, écrit-il, se donne elle-même pour la vérité la plus riche et la plus pauvre. Elle ne dit de ce qu'elle sait que ceci : *c'est* ; et sa vérité contient uniquement *l'être* de la chose qui l'occupe ; la conscience, de son côté, n'est dans cette certitude que comme pur Je, ou encore, Je ne suis que pur *celui-ci*, et l'objet, pareillement, n'y est que comme pur *ceci* [...] la conscience est Je, rien d'autre, un pur cet individu-ci [...] »⁹

Une autre source d'importance, pour le sujet qui nous préoccupe — mais je n'ai pas la prétention de les envisager toutes — est à chercher dans la tradition hindoue. On trouve en particulier une réflexion intéressante dans les enseignements oraux de Sri Ramana Maharsi, connu pour sa méthode d'investigation de soi centrée autour de la question « Qui suis-je ? » Ce *jivanmukta* (esprit libéré¹⁰) passait sa vie dans le silence et la contemplation, attirant à lui au fil du temps, et indépendamment de sa volonté, un nombre croissant de disciples. Il avait l'habitude de demander, à ceux qui lui adressaient des questions, « Qui est-ce qui me pose la question ? » Si on lui répondait par un nom, il lui demandait qui était derrière ce nom. Si on lui donnait une profession, il demandait qui était derrière cette profession. De fil en aiguille, il invitait son interlocuteur à chercher la réponse en lui, dans la pure conscience (*sat-chit-ânanda* ou l'Être-Conscience-Félicité). Dans un des rares entretiens à avoir été retranscrits, et traduit vers le français, il expose le cœur de sa doctrine :

Je ne suis pas ce corps physique, [...] ni les cinq organes de perception sensorielle, c'est-à-dire l'oreille, l'œil, la langue, le nez et la peau, et leurs fonctions correspondantes : l'ouïe, la vue,

le goût, l'odorat et le toucher. Je ne suis pas les cinq organes d'activité, c'est-à-dire les organes vocaux, les mains et les pieds, l'organe de procréation et l'anus, et leurs fonctions respectives : le langage, les mouvements du corps physique, la jouissance et l'excrétion. [...] Même l'esprit pensant, je ne le suis pas ; et pas non plus cet état d'ignorance dans lequel ne se trouvent que les impressions des objets, et non les objets eux-mêmes et leurs fonctions. [...] Après avoir rejeté tout ce qui a été mentionné ci-dessus comme n'étant « ni ceci ni cela », cette pure Conscience qui seule demeure - CELA je suis¹¹.

Il est important de ne pas confondre le Je avec la conscience vive en général. Bien que ces deux mots soient souvent employés de façon synonymique, la conscience n'est rien d'autre que le fait qu'il existe quelque chose de mental, plutôt que rien — que Je le sente ou non. C'est un pur CELA impersonnel et atemporel, que d'aucuns appellent l'Être, d'autres Dieu, d'autres encore le Mental ou l'Esprit Universel. Ces terminologies, à mes yeux, sont équivalentes, mais parler de « ma conscience » est, stricto sensu, un abus de langage assez commode pour dire Je.

Une analogie éclairante consisterait à comparer la conscience à l'espace géométrique, qui contient en lui tous les phénomènes matériels possibles, sans se réduire pour autant à l'ensemble d'entre eux. Bien que certaines personnes, après avoir disposé une clôture autour d'un lopin de terre, s'imaginent par là être devenues propriétaires de la superficie qu'elle comprend, ils ne sont pas pour autant — grand bien nous fasse ! — propriétaires de l'Espace avec une majuscule.

Si la conscience est bien universelle, la nature a voulu que nous ne traversons que cette petite portion clôturée d'expériences conscientes qui constitue le Je. Les limites du Je ne sont cependant pas manifestes dans notre expérience, pas plus que ne le seraient les limites visibles de l'Univers. Je peux me représenter qu'il existe une instance à l'origine de toutes mes expériences singulières, mais je ne peux pas appréhender cette unité dans sa totalité. Sartre résumait très bien cette idée dans la *Transcendance de l'Ego*, en exprimant en substance que la conscience est quelque chose qui se vit, non qui se contemple. C'est pourquoi Ramana Maharsia a bien raison de dire « CELA Je suis » et non pas « Je suis CELA ».

Toutes les clôtures que Je pense apercevoir à mon expérience consciente sont forcément limitatives. Ces choses auxquelles le Je est identifié à un instant donné, c'est le Moi. Si le suis donc le fait de sentir quelque chose plutôt que rien, le Moi quant à lui, est l'ensemble des sensations stables auxquelles Je m'identifie. J'opérerais même une distinction entre le Moi mondain, qui est ce à quoi Je m'identifie en tant qu'homme du monde ; et le Moi métaphysique, qui est ce à quoi je peux espérer m'identifier après avoir fait passer le concept du Moi mondain sur le billard de la critique

philosophique. Précisons cette pensée.

Le Je peut être vu comme une fenêtre sur le monde, une ouverture au travers de laquelle un contenu sensible se donne à contempler. Métaphoriquement, cette fenêtre est en déplacement constant. Le paysage qui défile face à lui n'est jamais le même. Et pour cause : s'il était immobile, ce paysage, Je n'aurais aucune possibilité de faire l'expérience d'être un Je distinct de ce qui est contemplé. Je ne serais, tout au plus, que le paysage lui-même. Et il en va ainsi de toutes choses.

Lorsque Je contemple une chaise, Je ne suis, à cet instant précis, rien d'autre que les impressions de la chaise : la vue de son dossier, l'odeur du pin, le toucher de ma main sur l'accoudoir et les pieds de bambou, tout cela constitue l'ensemble de mes expériences conscientes. C'est peut-être ce qui fait tenir à Huxley les mots suivants : « Je passai plusieurs minutes — ou étaient-ce des siècles ? — non pas simplement à contempler ces pieds en bambou, mais à les être effectivement » Il écrit cela sous l'effet de la mescaline, une drogue rituellement utilisée dans le chamanisme amérindien. Et d'ajouter plus loin : « ou plutôt à être moi-même en eux ; ou, pour être encore plus précis (car le "moi" n'était pas en cause dans cette affaire, non plus qu'en un certain sens, ils ne l'étaient, "eux") à être mon non-moi dans le non-moi de ce fauteuil¹². »

Du point de vue du vécu, le monde s'impose au Je dans son unité monolithique. Je ne peux pas immédiatement distinguer ma main de la chaise, pas plus que l'accoudoir du reste de la perception. Et ce n'est que parce que, à l'instant suivant, Je contemple autre chose, que surgit la possibilité de ne pas être uniquement cette chaise que Je vois ; mais bien le principe même de contemplation sous-jacent.

Il ne faut pas tomber de Charybde en Scylla : le contenu du Je, bien qu'il varie au cours du temps, ne peut changer du tout au tout. Si chaque impression du monde, à un instant donné, n'avait rien de commun avec les instants antérieurs, cela reviendrait, à chaque nouvelle perception, à devenir un être différent. Il en va ici comme dans une expérience de pensée donnée par Leibniz : dans ses *Essais de Théodicée*, il demande au lecteur s'il accepterait de renoncer à sa mémoire et d'être couronné roi aussitôt après ? L'homme du monde répondrait sans doute par l'affirmative. Mais à bien envisager les choses, cela ne reviendrait-il pas à mourir et à voir une autre personne, portant le même nom et la même apparence que soi, monter sur le trône ? Pour garantir la possibilité au Je de prendre conscience de lui-même, il faut que les contenus de la conscience changent au cours du temps, mais de telle sorte toutefois qu'un rapport de continuité s'établisse entre les sensations du monde à chaque instant. « Le *Je pense* signifie : je lie » disait Martin Heidegger¹³.

Les points fixes de la sensation

On en déduit qu'il doit exister, entre deux impressions du monde A et B, au moins un élément commun et au moins un élément disjoint. En l'absence de cette condition, A et B seront soit un seul et même instant statique, au sein duquel il n'y a pas de différences entre le Je et le monde ; soit deux mondes sans aucune relation de continuité, que contemplent deux Je disjoints. Ces deux Je s'abîment, chacun de son côté, dans la contemplation de leurs objets respectifs, effaçant tout également la distinction entre le sujet et le monde perçu.

Du reste, c'est bien parce que Je n'existe que dans le changement continu allant de A à B que, lorsque Je pense à un instant donné, l'instant n'est déjà plus là. L'intelligibilité du présent est contenue dans le futur immédiat. L'instant est donc une abstraction de la pensée sans laquelle le temps ne pourrait pas être pensé ; et quand on a le malheur de penser cet instant, il perd de son sens. C'est ce qui faisait dire à Saint-Augustin : « Si personne ne me demande ce qu'est le temps, je sais ce qu'il est ; et si on me le demande et que je veuille l'expliquer, je ne sais plus¹⁴. »

De cette discussion, on conclut qu'il doit exister des *points fixes* de la sensation. Non pas fixes au sens qu'ils seraient présents tous ensemble à chaque moment, mais en ce qu'ils assurent la liaison entre les instants deux à deux. Ces points fixes sont les repères du Je, son ancrage métaphysique. Ils ne sont pas forcément tous présents en permanence, mais certains d'entre eux peuvent devenir conscients lorsque l'attention, à force d'exercices divers, est portée sur eux. Ils sont donc ce qui accompagne le Je en continu.

Dans l'expérience vécue, les points fixes les plus évidents sont les sensations du corps : ce sont les jambes assurant le déplacement dans l'espace, les sens autorisant la perception, les mains qui se saisissent des objets, c'est la cinesthésie et la proprioception, qui permettent de sentir nos membres, leurs positions relatives, ainsi que l'action interne des organes. C'est le sang qui s'écoule dans nos veines, qui pulse dans nos tempes, qui bat tantôt lentement, tantôt comme le tambour dément des profondeurs, le souffle qui bruisse, doucement, du fond du ventre jusqu'à l'extrême pointe du nez, sans oublier la sensation de la peau dans les grands vents nocturnes, l'impression de la terre qui presse contre nos pieds, de la chair qui brûle sous la caresse du soleil, de la tête qui tourne sur son axe, des vêtements sur le dos, les coudes — le cou ! Tout ce cortège de sensations, qui accompagne partout le Je comme sa traîne royale, en lui assurant de régner sur le royaume de la conscience vive, sont des parties intégrantes du Moi.

Ces points fixes ne sont toutefois pas des objets du monde, mais des liaisons privilégiées existant entre les sensations. Car si je les envisage comme contenus de conscience, le corps, la tête, le cou, la sensation de mes bras, mes jambes, n'ont en eux rien de subjectif. Ce sont des objets de perception à part entière, qui ne se différencient pas, en tant que tels, de la table ou de

la tabatière. Ce sont des vécus, des manifestations de la conscience, et non plus le Je à part entière.

Le propre du corps n'est pas d'être d'une certaine forme, d'une certaine couleur, ni d'occuper un certain espace. Le propre du corps est d'assurer une liaison entre les sensations. Lorsque ma main gauche touche ma main droite, comme dans l'expérience imaginée par Merleau-Ponty, je fais l'expérience de sentir l'une par l'autre. Ce phénomène proprement magique ne se produit pas lorsque ma main caresse le pelage d'un animal. Dans ce cas, je sens ma main sur le poil dru, humide ou revêche de mon chat assoupi, mais je ne sens pas ce que ce poil ressent au contact de ma main. Le point fixe est donc le fait même de la sensation qui accompagne chaque instant de conscience. Plus précisément, on dira que le point fixe est la connexion privilégiée existant entre les différentes sensations.

Pour illustrer ce point, peut-être pourrait-on se demander ce qu'il se produirait si ma main devenait insensible ? Si mes yeux cessaient de voir ? Si mes épaules, mes bras, mes jambes, mon sexe cessaient d'être dotés de la capacité de sentir ? Si mon corps entier n'était plus qu'une pierre informe, une boue acre et sans saveur, tout aussi incapable de porter le monde en soi qu'une pâte à tarte tout juste prête à être mise au four ? L'état que l'on obtiendrait, en poussant jusqu'au bout cette expérience de pensée, serait équivalent à l'extinction du sujet. Ce serait une mort théorique d'autant plus brutale et énigmatique qu'il serait bien difficile de dire où elle commence. Tant qu'il reste ne serait-ce qu'une cellule pensante, un centimètre carré de peau sentiente, un cheveu susceptible d'être pincé sur la surface du crâne, le sujet est un Je. Mais dès que disparaît l'ultime point fixe, il n'est plus.

Si les points fixes étaient les parties d'un tout, au même titre que les cellules sont les parties d'un organe et que les organes sont les parties du corps, ou que les pensées sont des parties de la conscience, on pourrait les soustraire un à un par la pensée. On pourrait du moins envisager la possibilité de les soustraire. Or à procéder ainsi, on arrive à la conclusion absurde qu'en retirant quelque chose du tout de la sensation, la sensation elle-même disparaît dans sa totalité. Il faut donc soit (1) que les points fixes soient le tout en son entier, soit (2) qu'ils se manifestent, chacun individuellement, au sein du tout. Dans le premier cas de figure, cela revient à dire des points fixes qu'ils sont le Je. Le Je en effet est le fait même de penser, de sentir, d'être quelque chose plutôt que rien. Le Je est toute sensation, tout acte de conscience *possible*. Et dire des points fixes qu'ils sont le tout, c'est rendre caduque le concept même de point fixe. À ce compte-là, j'aurais mieux fait de ne jamais l'avoir introduit !

Il faut donc que les points fixes se manifestent dans le tout, sans être le tout lui-même. Comment cela est-il possible ? Parfois, les énigmes de philosophie trouvent leur réponse dans les plus simples plaisirs. Si un cuisinier vous montre un appétissant gâteau en précisant qu'il a mis dedans un ingrédient secret ; s'il ajoute, pour donner du grain à moudre à votre curiosité,

que cet ingrédient ne se trouve dans aucun morceau du gâteau en particulier, mais dans chacun d'entre eux, que direz-vous ? Qu'il s'agit du liant et du sel de la vie. Je suggère de voir les points fixes comme le lait, le sucre, les œufs qui, finement mélangés, occupent tout le volume du monde. Ils sont les liaisons existant entre chacun des contenus de conscience. Par « liaison » j'entends les associations privilégiées. Chaque pensée entraîne à sa suite un cortège d'autres pensées, d'images, de sons, d'odeurs, non pas d'après le résultat d'un jet de dé divin, mais en suivant les chemins longuement tracés par l'habitude.

Pour conclure, j'insiste sur le fait que nous nous faisons souvent une image du Moi, que j'appelle le Moi mondain, qui est aussi éloigné du Moi métaphysique que la lune est éloignée du Soleil. Le Moi métaphysique n'existe pas en tant que représentation, mais uniquement en tant qu'association des représentations. Il est l'histoire du Je. Non pas une histoire objective, tissée d'événements, de mots. Non pas le livre de la bibliothèque qu'on peut tâter des mains, ouvrir et consulter. Le Moi métaphysique est le filtre qui associe certaines idées les unes aux autres en fonction de ses expériences antérieures. C'est donc le prisme au travers duquel le monde se donne à contempler. C'est, dans sa nature profonde, un conditionnement. Les points fixes ne sont ni dans le corps, ni dans le cerveau, ni dans la sensation, mais dans le temps qui coule, comme un fleuve, d'une sensation à l'autre ; en répétant certains schémas. Pour cette raison, peut-être pourrait-on dire que la condition de l'être humain est d'être un subtil équilibre entre changement et éternel retour ?

S'il est vrai que le corps est plus proche d'un point fixe que tout autre objet du monde — il l'est certainement plus que ce livre ou ce gâteau que j'ai sous les yeux — le corps n'est pas en lui-même un point fixe. Certes il accompagne toujours chacune de mes expériences de conscience. Je ne peux pas sentir, penser, jouir et souffrir autrement qu'en présence de ce corps ; si bien que chercher à se l'arracher reviendrait à courir après sa queue. Mais ce qui fait dire que le corps est le Moi, ce n'est pas le corps en tant qu'objet du monde. C'est le fait qu'il existe une articulation entre les sensations. Le Moi métaphysique n'est pas cette idée mondaine du Moi que l'on se fait en se regardant dans la glace. Le Moi métaphysique est immatériel, invisible et indécomposable. C'est pourquoi on aurait tort de le chercher dans une partie du corps, à la façon dont Descartes ou Swedenborg s'entêtaient à extirper l'âme de la glande pinéale. L'idée d'un Moi objectif n'est qu'un mirage, un rêve, un songe-creux. C'est l'opium de la philosophie. Il est ce qui fait dire à Ernst Mach :

« L'habitude qui consiste à traiter le complexe inanalysé du Moi comme une entité indivise s'est souvent exprimée sur le plan scientifique de manière curieuse. On extirpe d'abord du corps le système nerveux, en tant que siège des sensations. Puis, dans

ce système nerveux, on privilégie le cerveau, cherchant en lui un point fixe qui assignerait un siège à l'âme elle-même, afin de sauver la soi-disant unité psychique. »

16.2.2 Le Soi et l'Inconscient

Il ne viendrait vraisemblablement à l'idée de personne, face à la question du « Qui suis-Je ? », de répondre « Je suis les points fixes de la conscience. », « Je suis le conditionnement de mes vécus » ou encore « Je suis le temps qui coule d'un vase vide vers un autre vase vide. ». Et quand bien même nous rétorquerions cela, ce ne serait pas une mince affaire que d'énumérer ces points fixes, car mis à part pour ce qui est des sensations du corps (et qui constituent d'ailleurs le b.a.-ba de la pratique méditative) nous n'avons pas forcément conscience des points fixes. Bien qu'ils soient au fondement du Moi, et que leur existence puisse être déduite *a priori*, ils ne sont pas toujours des objets concrets de la pensée.

Nous pouvons certes nous faire une idée de notre corps et de son fonctionnement, mais cette idée ne coïncidera que difficilement avec toute la complexe richesse de ce corps. Pire encore : il existe au sein de Moi bien des choses qui dépassent le corps visible et organique. Les traumatismes traversés au cours du développement psychologique, par exemple, laissent des stigmates dans le Moi, et dans le corps, qui se manifestent par des anxiétés, des douleurs somatiques ou des fantasmes. Ces singularités de notre monde intérieur constituent notre jardin secret. C'est sur ce terreau fertile que germent un ensemble de réactions caractéristiques dans lesquelles nous reconnaissons notre personnalité. Il y a donc bien, dans l'aquarium visible du Moi, les traces de monstres abyssopélagiques, des prédateurs antédiluviens, des anguilles fantomatiques, qui glissent, qui bruissent discrètement sur l'onde nue — et c'est tout l'art de la psychanalyse que d'observer cet invisible.

Lorsque Je prends, en tant que conscience, mon corps et ma personnalité pour objets spécifiques de mon attention, et que je leur applique les concepts qui sont à ma disposition pour les comprendre, je me forge une idée approximative du Moi. Cela se produit, par exemple, quand je passe un questionnaire de personnalité et que je m'écrie, avec dans les yeux la même ferveur que si Archimède eût redécouvert les lois de la flottaison : « Je suis ENFP ! » ou encore « J'ai un enneagramme de type 9 ! » C'est également ce qui est en jeu quand, à l'issue d'une épreuve sociale, telle qu'un concours, un entretien d'embauche ou la rencontre imminente d'un amoureux, je fais la démonstration d'être (ou de ne pas être) compétent, intelligent, beau, sensible, frustré ou désirable.

C'est bien d'ailleurs parce que le Moi n'est pas directement visible qu'il m'arrive souvent de douter de mes qualités et de me comparer à une norme imaginaire. Je cherche, dans le cadre d'une psychothérapie ou d'un concours,

l'attestation de particularités que je me soupçonne secrètement, mais dont je doute pourtant — avec au fond du cœur l'odieux sentiment d'imposture. En m'appuyant sur le regard de l'autre, je construis et je vérifie une certaine image de moi. Cette image du Moi est ce que j'appellerai le Soi. Cette image se fait au travers d'une définition. Il est obtenu à l'aide de mots et de symboles, de représentations, comme lorsque l'on dit « Je suis un tel, j'ai tel âge, j'occupe telle fonction ».

Il est d'ailleurs curieux que la première chose que l'on associe à un enfant tout juste sorti du ventre maternel, dont les yeux plissés par la venue au monde hésitent encore à se dégager de ce flou, de cette confusion, de cette paix qui était celle des limbes primordiaux ; il est fou de penser que l'on ne trouve rien d'autre à faire que de lui donner un nom. S'imagine-t-on que ces quelques syllabes ont le pouvoir de conjurer la condition métaphysique du sujet ? Croit-on qu'une syllabe référant à la mère de Dieu ou à un apôtre pourront sauver le Moi de cet état d'invisibilité auquel il semble condamné ?

« Le Moi ne peut, en aucun cas, être sauvé », disait Ernst Mach, mais peut-il être nommé ? Le Soi est un sortilège visant à rendre une unité objective abstraite, définitionnelle, pragmatique, à un Moi-messie, à un Moï-salvateur, qui a depuis longtemps renoncé à apparaître dans le monde conscient. Contrairement au Moi, le Soi dépend de toute une palette de jeux sociaux. Il cherche à confirmer sa définition de lui-même dans l'interaction avec autrui, particulièrement avec les personnes significatives institutionnellement — des enseignants, des juges, des figures morales — ou affectivement — les conjoints, les membres de la famille. Le Soi ne se contente pas de définir ce qu'il est dans l'espace objectif, au moyen d'un nom, d'un corps, d'un titre, d'une personnalité, d'une histoire. Il cherche également à se définir par le biais de ses idées.

Le Soi s'attache, comme à autant d'appendices inamputables, des opinions religieuses, politiques et philosophiques, des goûts, des aversions et toutes sortes de filtres limitatifs qui lui permettent de faire entrer le Moi dans une narration socialement validée. Ceci fait qu'échanger des opinions ou des conceptions personnelles peut être pénible, voire proprement dangereux, lorsque la conversation, sous des dehors bénins, met en jeu l'identité des interlocuteurs. Les réactions défensives ont souvent pour vocation de maintenir la construction, parfois fragile, à laquelle le Soi a opéré pour atteindre la certitude d'être un certain Moi tangible dans le monde — et non ce flux de sensations désorganisées d'un Je schizophrénique. Le Soi est donc tantôt extensible, tantôt rétractable, selon qu'il trouve un groupe au travers duquel se construire ; ou un groupe en opposition duquel se définir. L'un et l'autre vont généralement de pair.

La suite de cette thèse tentera de montrer, pas à pas, comment la conscience vive parvient à se faire une idée du Soi et pourquoi il demeurera toujours, en marge de cette idée, une part inconsciente et irréductible. Cela demandera, premièrement, de comprendre comment le Je peut prendre pour objet

les contenus sensibles qui le traversent de façon à ce que ces contenus deviennent des perceptions déterminées. Il s'agira ensuite de théoriser de quelle façon, et sous quelles conditions, ce percept peut devenir un concept, puis d'appliquer les conclusions aux points fixes de la perception. Pour l'instant, qu'il suffise de saisir que le Je est une dynamique produisant tantôt un sentiment intime, somatiquement déterminé et largement inconscient du Moi ; et tantôt une construction consciente, socialement et intellectuellement conditionnée, du Soi.

Le Moi et le Soi sont des concepts que nous manipulons, dans la vie de tous les jours, tout comme Monsieur Jourdain faisait de la prose : sans nous en rendre compte. Le Je, en revanche, demande un exercice philosophique de la pensée — ou la réalisation d'expériences mystiques et contemplatives — qui ne sont pas immédiates d'accès. On peut supposer que cette difficulté tient du fait que, à la naissance, le Je n'est pas encore un Je capable de conceptualiser les points fixes de la sensation et de s'en faire un concept stable, si bien que le jeune être humain vit dans un état de fusion prononcé avec le monde, dans lequel il n'est rien d'autre que ce qu'il voit, ce qu'il sent, ce qu'il ressent. Or au cours du développement neuropsychologique, il se produit un évènement étrange. Ce Je qui semblait premier, ce Je à partir duquel toute réalité consciente pouvait prendre son appui, s'efface au profit du Moi. Le jeune enfant, non seulement oublie que le monde n'existe qu'en lui, mais lorsqu'il s'en rappelle — ce qui ne manque pas d'arriver — il attribue la réalité de la conscience aux points fixes de la perception qu'il a sans cesse sous les yeux. Il ne dit pas comme Descartes : « Je pense, j'existe », mais il se demande « Le monde ne serait-il pas une illusion de *mon* cerveau ? » Ce faisant, il voit le cerveau comme une objectivité du monde responsable de la subjectivité de la conscience.

Comme il arrive souvent en philosophie, ce qui est premier au sens logique ne l'est pas toujours du point de vue chronologique. Le Je est premier sur le plan des concepts, puisqu'il est ce sans quoi il n'y aurait ni monde mental, ni aucune forme de Moi, mais ce Je ne se manifeste qu'au travers de contenus de conscience qui brouillent sa trace. C'est ainsi que l'étape métaphysique fondamentale, où le jeune être se fait tantôt sceptique, tantôt pyrrhonien, en niant tour à tour la réalité du monde et celle de ses perceptions, est également le lieu d'une perte originelle de la conscience vive au profit d'une conscience hybride : une conscience de Schrödinger ni morte, ni vivante, ni objective, ni subjective, ni spatiale, ni tout à fait mentale. Cette chimère métaphysique est responsable, sans aucun doute, des propositions d'après lesquelles « l'esprit est dans le cerveau », « l'âme est dans le corps » ou, dans un autre registre, « le monde n'existe que dans ma tête ».

Bien que fascinantes à première vue, et nécessaires à la résurrection bien sentie de la conscience vive, ces propositions métaphysiques, comme quoi « l'esprit est dans le cerveau », sont des épreuves superficielles du réel, des ébauches incomplètes du Je. Elles tombent dans les chiasmes métaphysiques

que j'ai dressés dans la section [12.3.2](#). Elles échouent notamment à prendre la mesure du fait que si le monde est envisagé de façon vraiment subjective, il ne sera rien qu'une pure expérience de conscience sans matérialité (et existant à ce titre "en-dehors" ou "en parallèle" du corps, du cerveau et de l'espace-temps objectif).

Cette pure expérience de conscience subjective, sitôt qu'elle se débarrasse des objectivités illusoires, n'en est pas moins renvoyée à une objectivité irréductible : le fait qu'il y ait quelque chose plutôt que rien. Pourquoi-Je ? D'où viens-Je ? Telles sont les questions qui demeurent. C'est ainsi que prendre conscience du Je ne nous délivre pas du monde objectif, pas plus qu'il n'élucide toutes les questions philosophiques, ni d'ailleurs certaines questions simples, comme savoir pourquoi le Moi meurt lorsque la tête est arrachée du corps ! Le *réel*, qu'il s'agisse du monde physique des atomes ou de l'Esprit en Général, continue d'être suggéré à la fois comme visée du Je (Je pense le monde) et comme sa condition de possibilité (Je viens du monde).

En dépit de l'oubli métaphysique du Je qui domine largement à l'âge adulte, et particulièrement dans nos sociétés marquées par l'extériorisation de la personnalité (qui veut que le Moi soit vu comme une chose du monde), l'être humain garde néanmoins, inscrite dans sa chair à l'encre indélébile, cette idée vague et incertaine d'après laquelle il y a une vérité première, un être ou une essence, dans l'ordre des phénomènes conscients (qui agirait derrière ceux-ci et leur donnerait, pour ainsi dire, une origine). C'est pourquoi l'adulte, même rigoureusement positiviste, reste tenté de croire qu'il est né d'un monde, ou d'un Tout, duquel son expérience actuelle tire son intelligibilité, dont il porte la marque et auquel il retournera. Cette conviction, aussi fermement qu'elle puisse être rejetée sur le plan intellectuel, ne se présente pas moins dans notre cœur à intervalle régulier — particulièrement lorsqu'on est mis face à la mort ou à la fatalité que le Moi n'a rien de durable.

C'est peut-être la raison pour laquelle, lorsque l'on explique à un enfant qu'il est né du ventre de ses parents et qu'il mourra un jour, celui-ci ne puisse se satisfaire pleinement de la réponse : « Mais qu'étais-Je avant cela ? » demande-t-il avec, dans ces yeux ronds, une évidence qui met les parents dans l'embarras, « et que serais-Je ensuite ? » La naissance et la mort du Je, qu'elles soient ou non réelles (c'est sans doute un indécidable), ne peuvent pas être conçues intellectuellement. Car penser que Je cesse de penser, ou sentir que Je cesse de sentir, n'en est pas moins une pensée ou une sensation. Il en va de la nature du Je d'être monde à tout instant, *d'être* tout simplement, sans avenir et sans passé. Il en va de lui comme de l'être dans le poème du Parménide :

*Mais comment ce qui est
pourrait-il être plus tard ? Comment aurait-il pu devenir ?
S'il est devenu, il n'est pas, pas plus que s'il doit être un jour.
Ainsi disparaissent la genèse et la mort inexplicables¹⁵.*

Cette conviction de fusion du Je avec le monde, bien que perdue depuis l'enfance, serait-elle à l'origine de la croyance d'une vie après la mort ? ou d'une naissance avant la naissance ? La conscience de l'être humain, toujours vive en dépit de son état d'agonie, continue d'être un Je en recherche de contenus de conscience desquels se nourrir. Et lorsque, par des circonstances qu'il reste à élucider, les limites fermement établies du Moi et du Soi viennent à tomber, même provisoirement, la condition première de Je comme totalité redevient visible. Ces expériences limites, qu'il s'agisse de passions amoureuses, de contemplations, de psychoses ou d'absorptions de substances, permettent de voir au-delà du Moi et de remonter progressivement, comme à contre-courant, jusqu'à l'origine première du mental.

Bien que la théorie de Tononi tente de dresser une analogie de structure entre les principes de fonctionnement de l'esprit et ceux de la matière, elle est victime de quatre faiblesses :

1. L'introspection est immédiate.
2. La conscience est à la fois unitaire et composée de partie.
3. La conscience peut être décrite de façon statique.
4. La conscience se contente de traiter de l'information issue de l'environnement.

Afin de surmonter ces impasses, je me fais l'avocat d'un cognitivisme méréologique. Cette nouvelle approche s'intéresse à la façon dont se forme des objets au sein de l'esprit humain ; en ce sens, c'est un cognitivisme, mais la perspective posée sur le réel est ontologiquement neutre.

Ce que je veux dire par là, c'est que les questions ne sont pas formulées en supposant que l'environnement extérieur agisse sur l'esprit. La question consiste plutôt à comprendre par quelle nécessité interne il est indispensable que se forme, au sein de la conscience, une distinction entre le registre de l'objectivité (le monde matériel) et de la subjectivité (l'expérience vécu).

Il découle de cette approche que la propriété essentielle de la conscience est d'être conscience de quelque chose qui change. J'appelle ce principe le principe de différenciation. Je développe en outre une méthodologie pour vérifier l'existence de ce principe, consistant à étudier ce qui se passe lorsque le cerveau est isolé de toutes sortes de stimulations (expérience de Cunningham).

À partir de cette méthodologie, j'esquisse brièvement la façon dont naissent, dans la conscience, une distinction entre le Je, le Moi et le Soi. Je tente en outre de décrire les lois dynamique régissant l'évolution psychique. Cette théorie embryonnaire fera l'objet de plus amples développement dans mon essai intitulé « La Conscience est-elle morte ? »

Conclusion

*Vous pensez que les robots ont
une âme ?*

Philip K. Dick

Une machine peut-elle penser ? Il est grand temps de résumer ce que les outils conceptuels que j'ai adoptés dans les précédentes parties peuvent nous enseigner à propos de la question capitale. Tout au long de cette thèse, je me suis appliqué à justifier, non sans diligence, qu'il existait trois couches à la pensée humaine : le niveau des percepts (sensation), le niveau des concepts intuitifs (perception) et le niveau des concepts rationnels (concepts abstraits).

Les concepts abstraits, ou rationnels, sont définis uniquement à l'aide de mots ou de nombres. Ce sont de purs signes syntaxiques. Ils ne capturent pas toujours une réalité sensible et expérimentable, mais peuvent être manipulés par le truchement du raisonnement logique. C'est ce niveau de pensée qui semble le plus proche du fonctionnement de l'ordinateur, mais c'est également la dimension la plus superficielle. Au moment de parler des théorèmes d'incomplétude de Gödel, j'ai indiqué que, dans une théorie axiomatisée un tant soit peu complexe, la majorité des propositions sont syntaxiquement indémonstrables : on ne peut pas dire si elles sont vraies ou fausses sur la seule base des axiomes. Les rares vérités qui succombent à la logique sont des points isolés, des îles solitaires, perdues au milieu d'un océan infini. La raison calculante, pour le dire autrement, est la peau qui recouvre le fruit, l'écorce qui dissimule la sève !

Le niveau des percepts quant à lui, est de l'ordre du sensible. La perception s'y présente comme un tout, un voisinage d'événements et de formes spatiales amalgamées. Il n'y a pas d'unités minimales de la perception, telles que des pixels ou des instants, qui servent de points de référence. En l'absence de ces briques élémentaires, l'esprit doit isoler des contrastes au sein de la totalité perceptive, en suivant un principe consistant à différencier et à assimiler les parties les unes aux autres. Des formes naissent, qui ne sont pas toujours conscientes. Ce sont tout au plus des sensations fugitives, parfois intenses, mais fortement liées au contexte de l'expérience, où il n'est pas

permis de distinguer aisément le sujet qui éprouve de ce qu'il éprouve. On ne peut pas isoler le percept et dire qu'il a une réalité en dehors de l'expérience sensible elle-même. Ce niveau du jugement que je décris est celui que tentent d'imiter les réseaux de neurones. Des corrélations statistiques sont établies de sorte à faire naître des catégories au sein des données sensibles. Bien que les résultats soient bons, je me suis efforcé de montrer que la façon dont ces algorithmes représentent les données présente des divergences par rapport à la perception humaine.

Le troisième niveau, situé entre les deux premiers, est celui des concepts intuitifs. Ici, l'esprit humain isole une forme perceptive et lui donne une réalité intelligible, consciente, exprimable linguistiquement. Dans ce processus s'établit une distinction entre le concept et la chose que ce concept désigne. Il est possible d'adopter un point de vue platonicien, dans quel cas le concept existe de façon idéale et les formes sensibles n'en sont que des expressions particulières. Mais il est également possible de prendre le point de vue empiriste, dans quel cas ce sont les formes sensibles qui sont premières et nos concepts ne font rien d'autre que de les synthétiser. Dans un cas comme dans l'autre, le concept est une réalité manipulable par la pensée, mais liée à un contenu sensible. Il n'est ni une pure abstraction syntaxique (comme dans le niveau de la Rationalité), ni une simple sensation singulière (comme dans la Sensation).

J'ai suggéré que les concepts intuitifs de ma théorie occupaient, vis-à-vis de la pensée kantienne, une position analogue aux jugements synthétiques *a priori* du type : « La somme de 5 et 7 fait 12 » ou « La somme des angles d'un triangle fait 180 degrés ». Ces derniers, en effet, ne se fondent ni sur la raison pure, ni sur l'expérience pure, mais sont des conséquences de la façon dont l'être humain structure les phénomènes. La grande idée de Kant, pour l'exprimer dans toute l'élégance de l'art, c'est que les données sensibles ne sont pas perçues telles qu'elles existent en soi, mais qu'elles prennent des formes qui dépendent de nos concepts et intuitions préalables. Ces derniers appartiennent au bagage inné de l'être humain. Lorsque nous pensons que l'Univers est un espace euclidien ou que le temps est une réalité objective, nous nous comportons à la façon d'un ingénieur zélé qui, après avoir apposé partout la marque de ses outils, pense que la nature a fait le monde avec les mêmes outils que lui.

Si la machine ne pense pas comme l'être humain, c'est parce qu'elle ne structure pas les phénomènes sensibles avec les mêmes outils que nous autres. Pour s'en rendre compte, voyons d'abord comment les choses se présentent dans la perception humaine, puis comparons-les à nos algorithmes. Au sein de notre esprit — je pense l'avoir dit plusieurs fois — il existe un rapport étroit entre le tout de l'espace et les parties de celui-ci. Les points de l'espace ne sont pas des données brutes ; ils tirent au contraire leur intelligibilité du lien entretenu avec les autres points du voisinage. Il y a ainsi une imbrication entre l'espace, pris dans son ensemble, et les figures qui se déploient en

son sein. C'est ce qui faisait dire à Kant que l'espace est le réceptacle des phénomènes, qu'il est ce sans quoi aucun corps ne peut être pensé.

Un phénomène analogue, quoique plus complexe à caractériser, se produit pour la perception humaine du temps. Les événements ne sont pas donnés sur une ligne chronologique allant du passé vers le futur. Le cerveau — auquel on impute bien des tours de magie — ne se contente pas de placer les stimuli à la file indienne. Aux yeux de l'esprit humain, il n'y a pas d'instant en soi, mais uniquement les changements intervenant sur des intervalles de temps. C'est pourquoi le temps subjectif dépend, pour nous, de la quantité de changement survenant dans nos vécus intérieurs : lorsque le paysage change, que nos pensées vagabondent, que nous rêvassons, que nous éprouvons des émotions fortes, le temps passe plus vite que lorsque nous restons dans une absolue fixité.

Or les algorithmes recourent à une toute autre représentation des données sensibles. Ils disposent pour commencer d'une horloge interne ; un certain mécanisme qui leur permet d'effectuer, à chaque unité de temps, exactement une opération numérique. De plus, la machine ne peut pas penser les rapports liant le tout et ses parties de façon interdépendante. Elle découpe l'information en unités de perception minimales : une photo est décomposée en pixels, un enregistrement audio est digitalisé, les mots sont décomposés en lettres et lues l'une après l'autre (là où l'être humain est attentif au mot dans son entier¹⁴). L'ontologie de la machine est un flatland, un monde plat, sans relief, où les savoirs s'obtiennent par juxtaposition de briques élémentaires. Pour le dire en une phrase lapidaire : les structures de perception humaine sont holistiques, tandis que les structures de la machine sont discrètes. Si cette idée que j'ai défendue s'avérait exacte, l'on détiendrait une différence de *fonctionnement* essentielle entre l'esprit humain et les algorithmes.

Le Mystérianisme perspectiviste

En analysant les problèmes de l'intelligence artificielle, et particulièrement le problème du jugement et de l'intuition, je ne me suis pas contenté d'un prisme de lecture kantien. J'ai au contraire critiqué les intuitions du temps et de l'espace introduites par Emmanuel Kant et je les ai remplacées par un nouveau couple structurel : l'intuition de l'Infini et l'intuition du Tout. J'ai ensuite cherché à caractériser formellement ces intuitions par trois fonctions mathématiques :

14. Une étude de l'utrivsineé de Crbindage a mrtnoé que l'on peut snas pbmlèore lrie un texte dnot les lrtets sont dnas le drrésode pour peu que la prierème et la dirnrée lttère de cauhqe mot rtenest à la bnone pacle. Ceci motnre que le careveu ne lit pas teuots les lteetrs mias pnred le mot comme un tuot. La pveure : avouez que vuos n'avez pas eu de mal à lire ce ttexe. (Pour un commentaire de ce paragraphe, voir THIERRY OPILLARD, *La Dégradation Cambridge : Même pas vrai ?*, Les Actes de Lecture n.102, juin 2008.

1. La complétion de suites infinies.
2. La résolution de problèmes indécidables.
3. La chance face à un problème NP-complet.

Une machine de Turing classique, comparable aux algorithmes que nous utilisons aujourd'hui, échoue à exécuter ces fonctions de façon satisfaisante. J'ai toutefois montré que si l'on permettait à cette machine de compter l'infini, en effectuant une infinité d'opération d'après certaines règles précises, alors ces trois fonctions sont réalisables. Le développement m'a conduit à montrer en quoi, précisément, la notion d'infini se cache derrière la résolution des problèmes indécidables, et en quoi la capacité à appréhender la totalité se cache derrière les problèmes NP-complets.

En prenant un peu de distance sur les résultats techniques, on pourrait dire que l'intuition du tout et l'intuition de l'infini ne semblent pas pouvoir être modélisées adéquatement par une machine de Turing classique. Ce qui est toutefois intéressant, c'est qu'il est tout aussi peu vraisemblable de pouvoir modéliser l'esprit humain avec une machine de Turing transfinie. La raison à cela est que notre cerveau, de taille finie, semble difficilement apte à réaliser des opérations de calculs illimitées.

Tout se passe donc comme s'il existait un mystère irréductible à l'endroit de l'esprit humain : les théories que nous avons pour décrire le cerveau sont insuffisantes pour capturer certaines propriétés essentielles de l'intuition. Inversement, les modèles idéalisés que nous construisons à propos de l'esprit ne sont pas modélisables en des termes physico-chimiques. J'ai donné à ce constat le nom de mystérianisme perspectiviste.

Au sein du mystérianisme perspectiviste, nous pouvons nous faire, à propos de l'esprit, deux types de représentations. Dans l'une d'elles, l'esprit est intelligible, mais non la façon dont cet esprit est créé à partir de la matière. Dans la seconde perspective, le fonctionnement matériel du cerveau est compréhensible, mais non la façon dont ce cerveau donne naissance à certains types précis de propriétés cognitives. J'ai discuté cette hypothèse en regard du développement récent des neurosciences et j'ai trouvé, auprès d'Edelman et de Tononi, des points de support crédibles à ma vision philosophique.

La Thèse forte et la Thèse faible

Il se dégage de mon travail deux thèses à propos du jugement de la machine, une thèse faible et une thèse forte. La thèse faible consiste à dire que la machine de Turing classique ne fonctionne pas comme l'être humain ; la thèse forte précise que les algorithmes actuels ne peuvent pas même imiter l'être humain. Je vais décrire ces deux thèses l'une après l'autre.

La thèse faible, dans sa substantifique moelle, stipule que l'intelligence artificielle ne fonctionne pas comme l'esprit humain en raison de son *onto-*

logie informationnelle. Cette notion d'ontologie renvoie à la façon dont les informations sont représentées, structurées et liées les unes aux autres ; ou autrement dit, à ce que c'est que *d'être* une machine. La thèse faible, en dépit de son nom, a de fortes implications : elle remet en question l'équation méthodologique implicite qui consiste à dire que l'esprit, c'est le cerveau ; et que le cerveau, c'est un réseau de neurones. Elle nous invite à réviser notre computationnalisme implicite, qui transpose, à l'étude de l'esprit humain, les propriétés des outils (ordinateur, logiciels...) que nous utilisons pour le modéliser :

Thèse 3 (Thèse faible). *L'esprit humain ne fonctionne pas comme les algorithmes actuels. On a tort de prendre les réseaux de neurones artificiels comme modèles du cerveau humain, et les modèles du cerveau humain comme des modèles de notre esprit.*

La thèse forte, quant à elle, stipule que non seulement l'esprit humain n'a pas la même ontologie informationnelle que la machine, mais qu'il n'est pas non plus *imitable* par une machine. Cette assertion revient à dire que l'esprit est analogue à une boîte noire dont on ignore le fonctionnement. Tout ce que nous pouvons faire, c'est créer des associations statistiques unissant l'information qui entre dans cette boîte à l'information sortante. J'ai suggéré que c'était exactement ce que faisaient les réseaux de neurones artificiels. Ils tentent d'imiter des fonctions qui sont (a) imparfaitement connues ou (b) incalculables.

La thèse forte n'implique pas forcément que certaines tâches, réalisables par l'esprit humain, sont inimitables dans l'absolu. Elle dit simplement que, pour toute machine donnée, il existe des tâches sur lesquelles cette machine ne sait pas imiter l'être humain. Cette formulation est plus prudente ; elle s'apparente, du reste, aux théorèmes d'incomplétude de Gödel (chapitre 5.2) et au théorème de Rice (5.3). Ces derniers ne prétendent pas que certaines propositions sont indémonstrables ou que certaines fonctions sont incalculables en soi. Ils affirment simplement qu'il n'existe pas de système d'axiomes ou de machine de Turing généralistes, capables de résoudre toutes les questions mathématiques. Le rêve de créer une intelligence artificielle forte, capable de totaliser le savoir humain et de résoudre tous les problèmes dont l'esprit se rend capable, voilà précisément ce à quoi s'oppose la version forte de ma thèse.

Notons que cette remarque que je fais à propos de la thèse forte s'applique assez bien aux exercices de QI artificiels conçus par François Chollet, dont j'ai parlé au chapitre .

Thèse 4 (Thèse forte). *L'esprit humain n'est pas entièrement imitable par les algorithmes actuels. Pour tout modèle algorithmique donné, il est possible d'exhiber une tâche sur laquelle le modèle échoue à imiter l'être humain.*

Je pense avoir donné des arguments pour suggérer que la thèse faible, à défaut d'avoir reçu une démonstration définitive, constitue en tout cas une objection sérieuse à l'endroit d'un réductionnisme *ad computatum*. Les analyses que j'ai faites de l'intuition humaine, via les illusions d'optique et les expériences neuroscientifiques de la chronostase, demandent à être contredites par les tenants de l'intelligence artificielle forte. Je pense que de ce débat, même s'il devait me donner tort, pourraient naître des idées intéressantes. Je n'aurai pas peur de m'installer au milieu de cette place publique — que l'on nomme communément le débat philosophique — avec une pancarte clamant haut et fort : *prove me wrong*.

Si la thèse faible peut faire l'objet d'arguments philosophiques de type spéculatif, je pense que la thèse forte, en revanche, relève d'une question empirique. Il n'est pas possible de déterminer, *a priori*, les performances de nos statistiques à *imiter* parfaitement une fonction imparfaitement connue. Et pour cause : la fonction étant inconnue, nous ne savons pas si elle est facile à imiter ou non. Et quand bien même nous détiendrions une méthode statistique imitant parfaitement la fonction de l'esprit sur toutes les instances connues, nous n'aurions pas pour autant de certitude concernant l'imitation des instances encore inconnues. Ce problème *d'extrapolation* peut être illustré au moyen d'une illustration simple.

Supposons que nous tentions d'imiter, à l'aide d'un réseau de neurones, la température à laquelle l'eau se change en vapeur en fonction de la pression. Pour la simplicité de l'exemple, je suppose que le réseau de neurones est un réseau à une seule couche qui fait de la régression statistique linéaire. C'est-à-dire qu'il se contente d'imiter les entrées et les sorties au moyen d'une fonction linéaire (une ligne droite). D'après la thermodynamique, la relation entre la température et la pression pour un liquide est linéaire jusqu'à un certain point singulier, qui est appelé point triple en raison de la coexistence des trois états de la matière : liquide, solide et gazeux. À partir du point triple, on observe un changement de pente (voir figure 16.2).

Avec l'entraînement approprié, notre réseau de neurones est susceptible de tracer une droite épousant la pente moyenne de la fonction, sans voir le changement de pente survenant à hauteur du point triple. Il en va en effet de la nature même de la régression statistique de « coller » les données au plus près, sans comprendre nécessairement les lois qui les gouvernent. Il va sans dire qu'un réseau de neurones plus intelligent peut comprendre la présence d'un changement de pente. L'être humain ne fait peut-être lui-même pas autre chose que des statistiques élaborées lorsqu'il tente de théoriser les relations physiques unissant différentes expériences. La difficulté de la science consiste à choisir soigneusement quelles variables mettre en relation au travers des protocoles de laboratoires les plus judicieusement choisis. Si l'on se contente de donner à la machine toutes les données connues, sans hypothèses ni lois préalables, des disparités risquent rapidement d'apparaître entre la prédiction du réseau de neurones et les observations réelles. Une critique

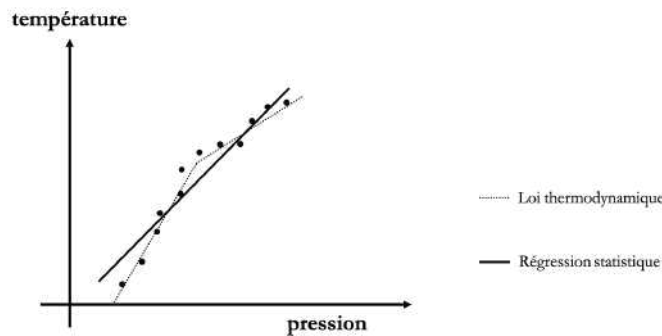


FIGURE 16.2 – La température à laquelle une matière se change en gaz dépend de la pression. On observe que la régression statistique ne peut pas saisir le changement de pente prévu par les lois de la thermodynamique.

bien sentie de l'intelligence artificielle nous reconduit, en dernier ressort, à une critique de la méthodologie scientifique.

Cet exemple simplifié que j'ai donné ne disqualifie pas la possibilité théorique d'imiter parfaitement les fonctions cognitives humaines, mais il souligne la présence de certains obstacles évidents. Si l'on me demandait de me prononcer en faveur ou en défaveur de la thèse forte que j'ai introduite en me mettant un fusil sur la tempe, je ne suis pas absolument sûr du pari que je prendrais. J'aurais tendance à défendre que l'imitation des fonctions cognitives humaines est rendue impraticable en l'état actuel de la science, à cause (1) des trop nombreuses inconnues qui planent encore sur le fonctionnement de l'esprit humain, (2) de l'ontologie discrète propre aux machines que nous utilisons actuellement et (3) des théorèmes de limitation intrinsèques de Gödel et de Turing. Il ne s'agit toutefois que d'une opinion aveugle, d'un vœu pieux, d'un cri époumoné poussé au cœur de la nuit.

Je m'estimerai satisfait si je suis parvenu à faire douter le lecteur à propos de certaines affirmations comme quoi « L'IA est une machine pensante », « le développement de l'IA va continuer de croître exponentiellement » ou que « le point singulier à partir duquel l'IA dépassera l'être humain est proche ». Avant de terminer cette thèse, j'aimerais encore esquisser brièvement certaines conséquences éthiques liées au jugement de l'intelligence artificielle.

ChatGPT : Une alchimie des savoirs

De la question de savoir si les machines ont la capacité de juger devrait dépendre l'orientation éthique des algorithmes. Est-il raisonnable de confier des décisions à des programmes qui sont inaptes à comprendre, autrement que par association statistique, les concepts intuitifs de justice, de liberté,

de dignité ? Des logiciels comme ChatGPT semblent être capables d'utiliser ces mots pour construire des discours, mais juge-t-on d'un homme politique à ses discours ? Ou à la vision qu'il véhicule sur la société future ?

ChatGPT a si bien défrayé les chroniques — et l'auto-apprentissage continue de s'exercer à une vitesse tellement impressionnante — qu'une question semble brûler sur toutes les lèvres des hommes et des femmes du monde informatique : l'intelligence artificielle ne représenterait-elle pas un danger ? Les performances des chatbots, lorsqu'il s'agit de donner une réponse vraie à un problème mathématique, ou une opinion argumentée à une discussion politique, sont sujettes aux plus hautes palinodies. Tantôt le programme donnera des informations correctes, tantôt il inventera, mais avec chaque fois la même allure de vraisemblance. Cet enfumeur professionnel, qui souffle le chaud et le froid, ne peut pas tenir le cap face à un expert attitré, mais il étonne tout de même par sa capacité à ne jamais chavirer.

D'aucuns pourraient penser que les réponses de la machine oscillent comme le vent et la pluie, avec toute l'imprévisibilité qu'on associe généralement à la météorologie. Dans la mesure toutefois où l'être humain lui-même — cet être humain moderne, pétri tout entier par les merveilles et les vicissitudes d'internet — n'est pas toujours garant de la vérité ; dans la mesure, du reste, où nos sociétés ont beau jeu de mettre sur un piédestal de savants imposteurs, on peut se demander si l'IA ne serait pas, ni plus ni moins, un être humain sujet à l'erreur et au génie, à l'exactitude comme à la falsification ? Elle représente en quelque sorte, en une seule figure, le George Psalmanazar de la géographie, le Grey Owl de la cause amérindienne, le Charles Ponzi de la finance, le Jean-Claude Romand de la médecine — ou plus proche de nous : le Stéphane Bourgoin des tueurs en série.

Une opinion, réservée et nuancée, sur les performances de ChatGPT, consisterait à distinguer la dimension *technique* de la dimension proprement *épistémologique*. Du point de vue technique, la prouesse des concepteurs est sans commune mesure, en ce qu'elle anticipe, de plusieurs années peut-être, les progrès attendus dans le domaine des chatbots. Il suffit de comparer les performances de ChatGPT-2, raciste, qui niait l'holocauste et voyait en Hitler un homme d'État plus à même qu'Obama de combattre le terrorisme, au ChatGPT-4 bien sage, trop sage peut-être, pour ne pas s'autocensurer à tire-larigot. L'IA d'aujourd'hui passe même le *Robot College Student Test* de Ben Goertzel, qui consiste à se faire passer pour un étudiant dans un collège et à réussir les examens. S'il est vrai que la machine ne soit pas présente physiquement dans les locaux, elle n'a guère de mal à satisfaire les évaluations. Elle réussit également, haut la main, l'examen d'entrée à l'école d'ingénieurs^[15]!

En dépit de ce triomphe technique, le bilan est plus nuancé sur le plan

15. AXEL LEGAY dans <https://www.rtb.be/article/chatgpt-lintelligence-artificielle-qui-inquiete-les-profs-elle-reussit-lexamen-dentree-pour-les-ingenieurs-11122991>

épistémologique. ChatGPT, loin de cet outil révolutionnaire dont on chante les louanges dionysiaques, ne serait, à cet égard-là, que la prolongation des moteurs de recherche existants. Entendons-nous là-dessus. Il est vrai, du point de vue d'un informaticien, que ChatGPT produit des réponses, non seulement plus rapides qu'une simple recherche Google, mais souvent plus précises aussi, à la question, par exemple, de savoir comment produire tel ou tel bout de code. Pour le programmeur du vingt-et-unième siècle, il est généralement plus efficace d'utiliser ChatGPT que d'introduire sa question sur un moteur de recherche ou de la poster sur un forum. Il n'en reste pas moins que, du point de vue strictement conceptuel, il n'y a pas de différence de principe entre ces deux opérations. Dans un cas comme dans l'autre, et nonobstant l'importante question de la traçabilité et de la propriété intellectuelle, il s'agit de rechercher une réponse au sein des informations agglomérées sur l'Internet. Le discours de ChatGPT se présente ainsi comme la synthèse, rondement mise en forme, des contenus fournis par les internautes — avec toutes les espèces d'erreurs ou de traits de génie que l'on pouvait déjà y trouver.

Cette prolongation du moteur de recherche nous expose ainsi, de façon accélérée, pour ne pas dire hyperbolique, aux promesses et aux dangers de l'Internet. Pour ne citer que quelques critiques parmi d'autres, il y a un risque, sur le plan de l'utilisateur individuel, que la démarche consistant à s'informer sur la toile, à terme, fasse moins intervenir le sens du jugement de celui qui s'y livre ; que les sources soient moins traçables ; que les justifications soient moins variées ou moins vérifiables ; que les opinions soient plus consensuelles.

Sur le plan systémique également, on pourrait craindre que ChatGPT soit utilisé pour influencer l'opinion. Loin d'être, à l'instar de Wikipédia, une plateforme gratuite visant, par le biais du consensus, à dégager des vérités objectives, on peut imaginer un modèle économique fondé sur la diffusion subtile de publicités ou de certaines opinions. Des conflits naîtraient, visant à visibiliser certaines réponses au travers du chatbot, un peu comme aujourd'hui les entreprises se trouvent au coude-à-coude pour le référencement de leurs pages Google. Il n'est sans doute pas inutile de rappeler, à cet égard, que Google a produit un algorithme permettant de reconnaître, avec plus ou moins de fiabilité, les contenus produits par ChatGPT afin de bannir, de son système de référencement, les pages qui ne contiendraient, en guise de contenu, que du texte généré de façon automatique.

Cet enjeu est crucial, si l'on se permet d'imaginer ce que deviendrait l'internet sitôt que l'immense majorité des pages répertoriées par les moteurs de recherche ne seraient plus que des lignes de ChatGPT. Non seulement il serait difficile de tracer d'où provient l'information mais, plus inquiétant encore, les algorithmes reposant sur le big data n'auraient plus à brasser, sur le net, que des produits issus des algorithmes eux-mêmes — avec tous les biais éthiques que cela pourrait produire. Ce serait l'équivalent d'un moulin

qui, pour produire de la farine, n'aurait que son propre blé à moudre !

Apocalypse artificieuse ou Jugement dernier ?

Un dernier danger sur lequel j'aimerais retenir l'attention est de nature politique : avec l'abondance des intelligences artificielles de type bottom up, on risque d'observer une plus haute normalisation des opinions et des comportements, ou un monopole plus concentré des vecteurs de l'information. Ce algorithmes se nourrissent en effet largement des données collectées sur l'internet. Or il est vrai que l'internet, loin de seulement nous ouvrir l'esprit à la diversité des sensibilités et des croyances constitutives de l'humanité, a aussi pour effet de radicaliser les opinions, de nous retrancher dans des sphères de croyances imperméables les unes aux autres. Quiconque ayant été sur la toile aura fait l'expérience, au moins une fois dans sa vie, de la haute redondance de l'information. À une question aussi innocente que : « qu'est-ce qu'une personne narcissique ? », vous trouverez une pléthore d'articles Doctissimo, Aufeminin.com et psychologies.com se paraphrasant les uns les autres. En dépit de leur apparente diversité, ils s'inspirent tous vaguement du DSM-5 et de la CIM-6, les deux manuels de référence en matière de troubles psychiatriques. Il peut même arriver, lorsque l'on se perd dans les hasards du net, que des paragraphes entiers se retrouvent copiés-collés d'un article à l'autre.

Cette expérience de duplication et de répétition des savoirs, du reste, est aussi celle que l'on éprouve (dans une certaine mesure) lorsque l'on fait un cursus universitaire. Quelle que soit la matière sur laquelle on jette son dévolu, l'on réalise, après plusieurs années d'études, que les ouvrages clefs de la discipline empruntent, peu ou prou, les mêmes sempiternels chemins de pensée. Vous demandez-vous ce qu'est « l'attention », ce qu'est la « dialectique » ou « l'intelligence » ? Vous trouverez dans l'état de l'art la tendance à citer les mêmes Socrate, Posner ou Binet, à opérer les mêmes divisions, les mêmes rapprochements, à formuler sans cesse les mêmes critiques. C'est peut-être la raison pour laquelle, aujourd'hui, il est devenu facile pour un professeur de corriger deux mille copies d'examen et d'encadrer dix thèses par an ? À la manière de Descartes qui, d'un simple coup d'œil à la table des matières, déterminait ce dont allait parler un livre, l'Homme du XXI^{ème} siècle, quant à lui, semble vouloir se contenter du titre de la dissertation.

On peut toutefois se demander si ce phénomène de normalisation des idées, et des comportements que ces idées induisent nécessairement, relève plus de la nature d'internet que de la somme statistique des milliards d'individus sur Terre ? Mais même à supposer que l'intelligence collective produise, pour ainsi dire naturellement, une convergence des chemins de pensée, il est difficile de nier que les outils de communication massifs, cristallisés par Google et Wikipédia — et sans lesquels l'intelligence collective ne se serait

pas aussi facilement exercée — induisent une uniformisation significative.

Il est aisé d'imaginer une société future dans laquelle l'imitation d'œuvres picturales ou textuelles sera rendue si évidente que les productions matérielles originales, et les compétences permettant leur création, seront de plus en plus valorisées. Plutôt que de concentrer l'innovation sur l'élaboration de nouvelles idées, et de breveter une invention sur base des principes théoriques qui en régissent le fonctionnement, il sera plus important de concevoir des produits concrets, réalisés sur des supports tangibles. Détenir une œuvre d'art en mains propres, à cet égard, sera considéré comme un luxe. Les modes d'élaboration de certificats d'authenticité, cachetés par le moyen de NFT (non-fungible token) ou de tout autre procédé numérique, constitueraient pour leur part un enjeu juridique et politique international.

En parallèle de cette première tendance, et en extrapolant le phénomène, on pourrait rêver — ou se méfier tout au contraire — d'une société dans laquelle les idées et les œuvres intellectuelles ne feraient plus l'objet d'une propriété personnelle ; mais appartiendraient au fond commun du patrimoine humain. Dans un tel monde, il serait prétentieux de s'arroger la paternité d'une idée, tout comme il est arrogant, aujourd'hui, de prétendre au savoir universel. Ce qui était, par le passé, la finalité même du travail intellectuel, passerait ainsi pour une excentricité. Il n'y aurait plus d'œuvre personnelle ni de travail individuel.

Cet univers civilisationnel que nous esquissons d'un coup de pinceau s'oppose toutefois à une doxa bien ancrée dans le monde d'aujourd'hui, et qu'Hannah Arendt a parfaitement illustrée : nous sommes (encore et toujours) dans une société du travail¹⁶. Et bien que le travail, asymptotiquement, en vienne à tomber sous le régime de la machine, c'est de ce travail, et du produit qui en découle, que l'individu justifie son statut social et économique. Il faudrait donc soit abandonner l'idée que le travail fasse la valeur de la personne, afin d'abroger partiellement le régime de la propriété intellectuelle, soit trouver une façon de réinventer le travail.

Car enfin, il est important de rappeler qu'un outil technique n'est jamais que la pointe immergée de l'iceberg. Derrière les modes de fonctionnement d'un algorithme se cache également, de façon parfois plus insidieuse, une certaine vision de la société. Cette « ruse de la raison numérique », pourrait-on dire en paraphrasant Hegel, est responsable de ce que, à l'insu de son concepteur, le logiciel va véhiculer malgré lui une transformation sociale. C'est ainsi que l'internet ne s'est pas contenté d'apporter des bases de données cyclopéennes à l'Homo Sapiens. Il n'est pas qu'un livre qui viendrait s'ajouter sur l'étagère, déjà pleine de poussières, du patrimoine universel. Les technologies numériques, en ce compris ChatGPT, participent bien plutôt d'une révision en profondeur de la distinction existante, jadis, entre l'espace public et l'espace privé ; entre l'opinion et le comportement ; entre le jugement et

16. Voir le prologue à la *Condition de l'Homme Moderne*.

la décision.

Dans le paradigme des Lumières, l'individu était comme maître et possesseur de son intimité intellectuelle. C'est d'ailleurs à l'ère moderne que le mot "intimité", qui désignait au Moyen-Âge exclusivement le lien affectif avec un tiers, est venu à désigner ce qu'il y a de personnel, de privé¹⁷. Or aujourd'hui, les algorithmes des réseaux sociaux, bien qu'ils ne puissent pas utiliser ces données pour tous les usages, en savent bien plus sur vos opinions politiques, votre orientation sexuelle, vos troubles psychologiques, votre empreinte digitale, que ne le soupçonnera sans doute jamais votre meilleur ami. Nous ne vendons plus uniquement notre force de travail (il est d'ailleurs important de remarquer que le labeur humainement nécessaire à la satisfaction des besoins vitaux diminue de façon de plus en plus substantielle) — nous vendons nos désirs, nos rêves, nos peurs, comme si nous étions devenus nous-même, au plus profond de notre subjectivité, une marchandise.

À ce titre, les législations comme le RGPD, plus que de protéger les données personnelles, sont un aveu d'impuissance. Elles constituent l'ultime soubresaut d'un régime institutionnel dépassé par la fulgurance du développement technologique. Croire que nos données nous appartiennent en vertu du consentement, c'est se créer une forme de réalité parallèle. Dans l'univers du droit, qui s'applique aux individus dotés de noms, de visages, de corps faits de chair et d'os, tout se passe comme si l'intimité faisait le fondement de la dignité humaine. Mais dans l'univers virtuel, constitué de chiffres, de statistiques, de modèles prédictifs, voilà longtemps que notre subjectivité est dépouillée de tout ce qu'elle avait de personnel. Le monde contemporain, plus que jamais, est schizophrénique. Si le mot « pouvoir » désigne à la fois la capacité descriptive à exercer une influence sur autrui et la capacité légale prescriptivement conférée par les normes morales, il est juste de dire que le « pouvoir » s'est scindé en deux entités antinomiques : le pouvoir réel potentiellement détenu sur les données personnelles dépasse de loin l'usage légalement conféré. C'est même en raison de cette séparation, toute théorique, que ce régime marchand d'illégalité ou d'allégalité, trouve son efficacité et sa légitimation.

Toujours est-il que ChatGPT semble soulever, dans le présent, une interrogation primordiale : d'après quels principes s'opèrent les capacités de synthèse de l'information déployées par l'intelligence artificielle et en quoi sont-elles susceptibles d'influencer le fonctionnement cognitif de l'être humain ? On a vu que le jugement chez Kant n'est rien d'autre que le moyen de ramener des impressions concrètes à l'unité objective de l'aperception. Si l'on fait l'économie du vocabulaire technique, et un tantinet jargonieux, le jugement désigne donc la capacité, pour le sujet pensant, de réunir des

17. Pour une généalogie des significations du mot « intimité » au travers des encyclopédies du XVII^{ème} siècle jusqu'à nos jours, se référer à de ISABELLE BERREBI-HOFFMANN, *Les métamorphoses de l'intime*, dans *Empan* 2010/1 (n°77), pp.13-17

images, des mots, des phrases, des pensées ou des sensations sous une perception englobante. Il n'est donc pas réductible à des définitions rationnelles ni à des associations statistiques. Or en l'état, on est en droit de douter que l'IA soit en mesure de produire un sens subjectif à partir de la série de mots objectifs, et adéquatement liés les uns aux autres, qu'elle est en mesure d'aligner. Veut-on réellement confier la gouvernance de nos sociétés et de nos comportements à un juge sans yeux, sans oreilles, sans âme ? Nous regarderions un tel juge avec, dans la bouche, les célèbres mots d'Albert Einstein :

*Dans nos efforts pour comprendre [l'Intelligence Artificielle], nous sommes un peu comme un homme essayant de saisir le mécanisme d'une montre fermée. Il voit le visage et les aiguilles en mouvement, entend même son tic-tac, mais il n'a aucun moyen d'ouvrir le boîtier*¹⁸.

18. Je me suis permis de changer le terme « réalité », figurant dans la citation originale, par le terme « intelligence artificielle » (que j'ai mis entre crochets). Voir ALBERT EINSTEIN, LEOPOLD INFELD, *The Evolution of Physics*, 1938.

Bibliographie

- [1] ALEKSANDER I, GAMEZ D. *Informational Theories of Consciousness : A Review and Extension*, in Hernández C, Sanz R, Gomez J, Smith LS, Hussain A, Chella A, Aleksander I (eds) *From brains to systems : brain-inspired cognitive systems*, 2010, Springer, Berlin, pp 139–147.
- [2] ANGWIN JULIA, LARSON JEFF, MATTU SURYA AND KIRCHNER LAUREN, *Machine Bias Risk Assessments in Criminal Sentencing*, ProPublica, May 23, 2016 (<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>).
- [3] APPEL KENNETH & WOLFGANG HAKEN, *Les progrès des mathématiques*, Éditions Belin, 1981, 8-19 p. (ISBN 2902918143), « La solution du problème des quatre couleurs ».
- [4] ARENDT HANNAH, *Condition de l'Homme Moderne*, Le Livre de Poche, 2020.
- [5] ARRÁEZ-AYBAR, LUIS-A et al., *Thomas Willis, a Pioneer in Translational Research in Anatomy (on the 350th Anniversary of Cerebri Anatome)*, Journal of Anatomy, 226 (3) : 289–300.
- [6] ARISTOTE, *Métaphysique*, texte traduit par Christian Rutten et Annick Stevens, présenté et annoté par Annick Stevens, dans : Aristote, Œuvres, tome I, Paris, Gallimard, La Pléiade, 2014.
- [7] ARISTOTE, *De la Génération et de la corruption*, trad. Jules Barthélemy-Saint-Hilaire, Ladrangue, Paris, 1866.
- [8] ARISTOTE, *On Ideas : Aristotle's Criticism of Plato's Theory of Forms* (ΠΕΡΙ ΙΔΕΩΝ), commented by Gail Fine, Oxford University Press, 1993.
- [9] BENGT ASPVALL, MICHAEL F. PLASS ET ROBERT E. TARJAN, *A linear-time Algorithm for Testing the Truth of Certain Quantified Boolean Formulas*, Information Processing Letters, vol. 8, no 3, 1979, p. 121-123.
- [10] ATLAN HENRI, *Spinoza and la Biologie actuelle*, Odile Jacob, 2018.
- [11] ATLAN HENRI, *Les étincelles de hasard Connaissances spermatiques*, La Librairie du XXIe siècle, Col. Sciences humaines, 1999.

- [12] ATMANSPACHER HARALD, *Le monisme à double aspect selon Pauli et Jung*, 2014-1, N. 3, pp. 105-135
- [13] AUCHATRAIRE ANNE, *Le cogito cartésien et la Critique de la raison pure*, in *Figures du Transcendantale*, dir. CRAMPE-CASNABET, MICHÈLE, ENS Éditions, Lyon, 1996.
- [14] BAARS B. J., *A cognitive theory of consciousness*, Cambridge University Press, Cambridge, 1988.
BARJAMOVIC GOJKO, GONZALEZ PATRICIA JURADO, GRAHAM CHELSEA A., LASSEN AGNETE W., NASRALLAH NAWAL, AND SÖRENSEN PIA M., *Food in Ancient Mesopotamia : Cooking the Yale Babylonian Culinary Recipes*, in *Ancient Mesopotamia Speaks : Highlights of the Yale Babylonian Collection*, distribué par Yale University Press for the Yale Peabody Museum of Natural History, avril 2019.
- [15] BAROT EMMANUEL, *En quoi la crise des fondements des mathématiques est-elle terminée ?*, *Philosophia Scientiæ* [En ligne], 9-2 | 2005, mis en ligne le 15 juin 2011, consulté le 17 octobre 2024. <http://journals.openedition.org/philosophiascientiae/517>.
- [16] BAUDELAIRE CHARLES, *L'Homme et la Mer* dans *Les Fleurs du mal*, Paris, Michel Lévy frères, 3e éd., 1868.
- [17] BÉLANGER MATHIEU, *La méthode axiomatique durant la crise des fondements* in *La crise des fondements : quelle crise ?*, Les Cahiers d'Ithaque, séminaire du 7 au 9 octobre 2011.
- [18] BENER PATRICIA, *From Novice to Expert : Excellence and Power in Clinical Nursing Practices*, Addison-Wesley, 1984.
- [19] BENNETT DANIEL M., DENNET D., HACKER P., SEARLE J., *Neuroscience and Philosophy : Brain, Mind, and Language*, Columbia University Press, 2003.
- [20] BERREBI-HOFFMANN ISABELLE, *Les métamorphoses de l'intime*, dans *Empan* 2010/1 (n°77), pp.13-17.
- [21] BETTS LR, TAYLOR CP, SEKULER AB, BENNETT PJ. (2005) *Aging Reduces Center-Surround Antagonism in Visual Motion Processing*, *Neuron*. 2005 Feb 3 ;45(3) :361-6. doi : 10.1016/j.neuron.2004.12.041.
- [22] BLAKEMORE COLIN, HORACE BARLOW & WESTON-SMITH MIRANDA, *Images and Understanding : Thoughts about Images : Ideas about*, Cambridge University Press, 1990.
- [23] BLAKEMORE COLIN, *Mechanics of The Mind*, Touchstones, Londres 1997.
- [24] BLÅSJÖ VIKTOR, *First proofs : Thales and the beginnings of geometry*, Intellectual Mathematics, Podcast S02 : History of mathematics, 16 mai 2020, <https://intellectualmathematics.com/blog/first-proofs-thales-and-the-beginnings-of-geometry/> (consulté le 6 novembre 2024).

- [25] BLOCK NED, A Theory of Consciousness, in Block, Ned, Flanagan, Owen & Guzeldere, Guven, 1997 : *The Nature of Consciousness : Philosophical Debates*, Cambridge : MIT Press.
- [26] BOHM DAVID, *A New Theory of the Relationship of Mind and Matter*, Philosophical Psychology, 1990, 3(2-3), pp. 271–286.
- [27] EMIL BOREL, *Leçon sur la théorie des fonctions*, University of California Libraries, 1898.
- [28] BORGES, *Funes ou la mémoire*.
- [29] CALUDE CRISTIAN S. et JUERGENSEN HELMUT, *Is Complexity a Source of Incompleteness ?*, Arxiv, 5 Nov 2004, [arXiv:math/0408144](https://arxiv.org/abs/math/0408144).
- [30] CANTOR GEORG, *Sur les fondements de la théorie des ensembles transfinis*, 1895-1897.
- [31] CARREY SUSAN, *The Origin of Concepts*, Oxford University Press, 2009.
- [32] DAVID J. CHALMERS, *The Conscious Mind*, Oxford University Press, Inc., 1996.
- [33] CHOLLET FRANÇOIS, *On the Measure of Intelligence*, ArXiv, 5 novembre 2019.
- [34] CHOLLET FRANÇOIS, *ARC-AGI*, consulté le 17 octobre 2024, <https://arcprize.org/arc>
- [35] CHURCHLAND PAUL, *Eliminative Materialism and the Propositional Attitudes*, Journal of Philosophy, vol. 78, no. 2, 1981, pp. 67-90.
- [36] CHURCHLAND PATRICIA, *Neurophilosophy : Toward a Unified Science of the Mind/Brain*, Cambridge (Mass.), MIT Press, 1986.
- [37] ANDY CLARK, *Natural-born Cyborgs : Minds, Technologies, and the Future of Human Intelligence*, ed. Oxford University Press, USA, 2004.
- [38] COMTE AUGUSTE, *Système de politique positive*, Hermann, 2022.
- [39] COOPER J., *Plato : Complete Works*, Hackett ; 1997.
- [40] , *Hypercomputation*, in Minds and Machines, 12, 461–502 (2002). [https ://doi.org/10.1023/A :1021105915386](https://doi.org/10.1023/A:1021105915386).
- [41] COTNOIR AARON J., VARZI ACHILLE, (2021) *Mereology*, Oxford, OUP., ISBN 978-0-19-874900-4, DOI 10.1093/OSO/9780198749004.001.0001.
- [42] CRICK F., *Astonishing Hypothesis : The Scientific Search for The Soul*, Cambridge University Press, Cambridge 1995.
- [43] DAMASIO R. ANTONIO, *L'Erreur de Descartes*, Odile Jacob, 1997.
- [44] DAMASIO R. ANTONIO, *Spinoza avait raison. Le cerveau de la tristesse, de la joie et des émotions*, Odile Jacob, 2003.
- [45] DARRIGOL OLIVIER, DUPLANTIER BERTRAND, RAIMOND JEAN-MICHEL, RIVASSEAU VINCENT, *Niels Bohr, 1913-2013*, Poincaré Seminar 2013 in *Progress in Mathematical Physics* (PMP), volume 68, 2016.

- [46] DEHAENE ET AL., *The Mental Representation of Parity and Number Magnitude*, J. Exp. Psychol., 122, 371–396, 1993.
- [47] STANISLAS DEHAENE, *L'accès à la conscience* Discours tenu au Collège de France en 2009-2010. Vidéo YouTube (consulté le 17 octobre 2024) : https://www.youtube.com/watch?v=Zt_fMVnbDYE.
- [48] DEHAENE STANISLAS, CHANGEUX JEAN-PIERRE et NACCACHE LIONEL, *The Global Neuronal Workspace Model of Conscious Access : From Neuronal Architectures to Clinical Applications in Characterizing Consciousness : From Cognition to the Clinic ?*, Researches and perspectives in neurosciences, Springer, 2011.
- [49] DELEUZE LUC, *Lettres et autres textes*, édition préparée par David Lapoujade, Paris, Les Éditions de Minuit, 2015.
- [50] R. DESCARTES (1649) *Les Passions de l'âme*. France : n.p., 1728.
- [51] DESCARTES, *Discours de la méthode*, annoté par LAURENCE RENAULT, Paris, Flammarion, coll. « GF », 2016.
- [52] DEWALQ ARNAUD, *Deux théories de l'analyse psychique : Wundt et Brentano in Vers une philosophie scientifique : Le programme de Brentano*, Paris, Demopolis, 2014.
- [53] PHILIPPE K. DICK, *Les androïdes rêvent-ils de moutons électriques ?*, Éditions J'ai Lu, 16 février 2022.
- [54] DREYFUS HUBERT L., *Mind Over Machine*, Free Press, 12 septembre 1988.
- [55] DREYFUS HUBERT L., *What Computer Still Can't Do ?*, Harper & Row Publishers, New York, 1972.
- [56] EARL B., *The Biological Function of Consciousness*, Front Psychol 5 :1–18, 2014.
- [57] EDELMAN GERALD M., *Antibody Structure and Molecular Immunology*, Nobel Lecture, December 12, 1972/
- [58] *The Remembered Present : A Biological Theory of Consciousness*, 1989.
- [59] EDELMAN GERALD, *Bright Air, Brilliant Fire : on the Matter of the Mind*, Penguin, Harmondsworth 1997, pp.109.
- [60] EINSTEIN ALBERT, INFELD LEOPOLD, *The Evolution of Physics*, 1938.
- [61] EDELMAN GERALD, TONONI GIULIO, *Consciousness and Complexity*, décembre 1998, 282(5395) :1846-51.
- [62] EDELMAN GERALD , TONONI GIULIO, *Comment la matière devient conscience ?*, Odile Jacob, 1er mai 2000.
- [63] EDELMAN GERALD, *Naturalizing Consciousness : A Theoretical Framework*, PNAS, avril 2003, 100 (9) 5520-5524.
- [64] EDELMAN GERALD, *Plus vaste est le ciel : Une nouvelle théorie générale du cerveau*, Odile Jacob, Paris, 2004.

- [65] EDELMAN GERALD, *Second Nature : Brain Science and Human Knowledge*, Yale University Press, 2006.
- [66] (1964) *Eliza*, program, <https://www.masswerk.at/elizabot/eliza.html> (consulté le 17 octobre 2024).
- [67] EUCLIDE, *Éléments*, trad. François Peyrard, *Les œuvres d'Euclide, en grec, en latin et en français : d'après un manuscrit très-ancien qui était resté inconnu jusqu'à nos jours*, vol. 1-3, Paris, M. Patris, 1814, 1816, 1818.
- [68] YANG, EUNICE ; TADIN, DUJE ; GLASSER, DAVIS M. ; HONG, SANG WOOK ; BLAKE, RANDOLPH ; PARK, SOHEE (2012-11-15). *Visual Context Processing in Schizophrenia*. Clinical Psychological Science. 1 (1) : 5–15
- [69] EVAN PETER W., *A study of Time in Modern Physics*, Doctoral Dissertation, The University of Sydney, Philpaper, 2011.
- [70] FALLON JAMES, *The Psychopath Inside : A Neuroscientist's Personal Journey Into the Dark Side of the Brain*, Hardcover, October 31, 2013.
- [71] Fingelkurts AA, Fingelkurts AA, *Information Flow in the Brain : Ordered Sequences of Metastable States*, Information 8(1) :22, 2017.
- [72] FNAIECH NADER, FNAIECH FARHAT, JERVIS B.W. & M. CHERIET, *The Combined Statistical Stepwise and Iterative Neural Network Pruning Algorithm*, November 2009, Intelligent Automaton & Soft Computing 15(4) : 573-589.
FOUCAULT [?], *L'Herméneutique du Sujet*, Paris, Hautes Études/Gallimard/Seuil, 2001.
- [73] FREGE GOTTLOB, *Kleine Schriften*, I. Angelleli (ed.), Hidesheim : Georg Olms, 1967.
- [74] FREGE GOTTLOB, *Les Fondements de l'arithmétique*, trad. CLAUDE IMBERT, L'ordre philosophique, Seuil, 1969.
- [75] FREGE GOTTLOB, *Briefwechsel mit D. Hilbert, E. Husserl, B. Russell, sowie ausgewählte Einzelbriefe Freges*, hrsg. von G. Gabriel, Fr. Kambartel, Chr. Thiel, Hamburg, Felix Meiner, 1980, p. 9 (trad. fr. modif. in Logique... , p. 223).
- [76] FREUD SIGMUND, *Esquisse d'une psychologie*, trad. de l'allemand, Toulouse, Erès Collection Scripta, 2011.
- [77] FRIEDBERG, R.M., *A Learning Machine : Part I*, IBM Journal of Research and Development, vol. 2, Issue 1, January 1958.
- [78] GARDNER HENDRIE, *Oral History of Richard Greenblatt*, interview, 2005 Computer History Museum, CHM Reference number : X3131.2005.
- [79] GREGORY R.L. AND GOMBRICH E.H., *The Confounded Eye* in eds. Illusion in Nature and Art, Duckworth, London, 1973.

- [80] GOLD I., STOLJAR D., *A Neuron Doctrine in the Philosophy of Neuroscience*, Behavioral and Brain Sciences, 22(5), 809-30 ; discussion 831-69.
- [81] HABERMAS JÜRGEN, Théorie de l'agir communicationnel, titre original : Theorie des kommunikativen Handelns, 2 t. (1981); réédition : Paris, Fayard, 1987.
- [82] HAGGARD P, CLARK S, KALOGERAS J., *Voluntary action and conscious awareness*. Nat Neurosci. 2002 ;5 :382–385.
- [83] JOEL DAVID HAMKINS & ANDY LEWIS, *Infinite time Turing machines*, in Cambridge University Press : 12 March 2014
- [84] BABAK HASSIBI, DAVID G. STORK, *Second Order Derivatives for Network Pruning : Optimal Brain Surgeon*, Ricoh California Research Center, Stanford, CA 94305.
- [85] GEORG WILHELM FRIEDRICH, *Phénoménologie de l'esprit de Hegel*, Poche, 2012.
- [86] HEIDEGGER MARTIN, *Être et temps*, trad. Emmanuel Martineau, Édition numérique hors commerce, http://t.m.p.free.fr/textes/Heidegger_etre_et_temps.pdf (consulté le 10 novembre 2024).
- [87] HODINOTT-HILL IONA, THILO KAI V., COWEY ALAN, WALSH VINCENT, *Auditory Chronostasis : Hanging on the Telephone*, 15 octobre 2002.
HOWDYSHELL STANFORD, *On Realism and the Pessimistic Meta-Induction*, Kent State University, Critical Reflections, Essay on Significance and Critical Reflections.
- [88] HOFSTADTER, *Gödel, Escher, Bach : An Eternal Golden Braid*, Basic Books, janvier 1980.
- [89] HUME DAVID, *An Enquiry concerning Human Understanding*, Londres, 1739-1740.
- [90] JOHN F. IRELAND, MARVIN W. KAHN, *How Fair Is the Culture I.Q. Test ?*, Sage Journal, V.25-1, April 1979.
- [91] JAMES WILLIAM, *The Principles of Psychology*, Holt, New York, 1890.
- [92] JAYNES JULIAN, *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Houghton Mifflin, 1976.
- [93] JESHION ROBIN, *Frege's Notions of Self-Evidence*, Mind, Oxford University Press, Vol. 110, No. 440 (Oct., 2001), pp. 937-976.
- [94] JOHNSON AYSJA, VONG WAI KEEN, LAKE BRENDEN M. , & GURECKIS TODD M., *Fast and Flexible : Human Program Induction in Abstract Reasoning Task*, arXiv, 10 mars 2021.
- [95] JONKISZ J., *Consciousness : Individuated Information in Action*, Front Psychol 6 :1–17, 2015.

- [96] JONKISZ J. *Subjectivity : a Case of Biological Individuation and an Adaptive Response to Informational Overflow*, Front Psychol 7 :1–6, 2016.
- [97] JUNG CARL GUSTAV, PAULI WOLFGANG, *Naturerklärung und Psyche*, Zurich Rascher, 1952.
- [98] KAHNEMAN (2011), *Système 1 et Système 2 : les deux vitesses de la pensée*, Macmillan. ISBN 978-1-4299-6935-2. Retrieved April 8, 2012.
- [99] FRANÇOIS KAMMERER, *Conscience et matière*, Éditions matériologiques, Collection Sciences & philosophie, 2019.
- [100] KANDEL ERIC R., *La Biologie moléculaire de la mémoire : un dialogue entre gènes et synapses*, M/S : médecine sciences, 19(5), 625–633.
- [101] KANT EMMANUEL, *Critique de la Raison pure*, trad. de BARNI JULES, Édition Germer-Baillière, 1869.
- [102] KANT EMMANUEL, *Critique de la raison pure*, trad. TISSOT JOSEPH, Librairie philosophique de Ladrangue, édition originale 1836, édition révisée 1845 (Tome I) et 1869 (Tome II).
- [103] KANT EMMANUEL, *Fondements de la Métaphysique des mœurs*, 1785, Traduction de V. Delbos (1862-1916), Éditions Les Échos du Maquis, v1, juin 2013.
- [104] KANT EMMANUEL, *Prolégomènes à toute métaphysique future*, trad. GUILLERMIT LOUIS, préf. VUILLEMIN JULES, Paris, Vrin, coll. « Bibliothèque des textes philosophiques », 2001.
- [105] KAYE R., *The Minesweeper is NP-complete*, The Mathematical Intelligencer, Volume 22, number 2, 2000, pp. 9–15.
- [106] KAITARO TIMO, LOKHORST G.J.C., *Descartes and the Pineal Gland*, in Standord Encyclopedia, first published Mon Apr 25, 2005 ; substantive revision Wed Nov 5, 2008.
- [107] KHAYYAM OMAR, *Commentary on the Difficulties Concerning the Postulates of Euclid's Elements*, décembre 1077.
- [108] JAEGWON KIM, *Making Sens of Emergence, Philosophical Studies : An International Journal for Philosophy in the Analytic Tradition*, Vol. 95, No. 1/2, Reduction and Emergence, The Thirty-Third Oberlin Colloquium in Philosophy, (Aug., 1999), Springer, pp. 3-36.
- [109] JAEGWON KIM, *La Survenance et l'esprit*, Broché – Grand livre, 9 janvier 2009.
- [110] KINZEL KATHERINA, *Neo-Kantian Conceptualism : Between Scientific Experience and Everyday Perception*, British Journal for the History of Philosophy, 08 Aug 2023.
- [111] MORRIS KLINE, *Les Fondements des mathématiques, La Recherche*, mars 1975, num. 54, vol. 6, pp. 200-208.

- [112] KNUTH DONALD E., *The Art of Computer Programming : Fundamental Algorithms*, Addison-Wesley, 1997.
- [113] LADRIÈRE JEAN, *Les Limitations internes des formalismes : Étude sur la signification du théorème de Gödel et des théorèmes apparentés dans la théorie des fondements des mathématiques*, Ed. Jacques Gabay, 1992.
- [114] YANN LE CUN, DENKER JOHN S. et SOLLA SARA A. , *Optimal Brain Damage*, AT&T Bell Laboratories, Holmdel, N. J. 07733.
- [115] LEGAY AXEL, *ChatGPT, l'intelligence artificielle qui inquiète les profs : "Elle réussit l'examen d'entrée pour les ingénieurs!"*, <https://www.rtbef.be/article/chatgpt-lintelligence-artificielle-qui-inquiete-les-profs-elle-reussit-lexamen> (consulté le 5 novembre 2014).
- [116] LEMBEZAT CAROLE, *Google suspend un ingénieur qui assure que son IA est "consciente"*, Courrier international, 23 juillet 2022, <https://www.courrierinternational.com/article/polemique-google-suspend-un-ingenieur-qui-assure-que-son-ia-est-consciente>.
- [117] LEPAGE FRANÇOIS et FRADET KARINE, *La crise des fondements : quelle crise ?*, Les Cahiers d'Ithaque, séminaire du 7 au 9 octobre 2011.
- [118] LESLIE G. VALIANT, *Knowledge Infusion : In Pursuit of Robustness in Artificial Intelligence*, Foundations of Software Technology and Theoretical Computer Science (Bangalore) 2008 (conférence), janvier 2008.
- [119] LINSKY BERNARD, *Ernst Schroeder and Zermelo's Anticipation of Russell's Paradox in La crise des fondements : quelle crise ?*, Les Cahiers d'Ithaque, séminaire du 7 au 9 octobre 2011.
- [120] LOBATCHEVSKY NIKOLAI IVANOVICH, *Nouveaux principes de géométrie avec une théorie complète des parallèles*, Kazan, 1835-1838.
- [121] LOCKE, *Essai sur l'Entendement humain*, Traduction par COSTE PIERRE, Pierre Mortier, 1735 (p. Titre-321).
- [122] LUKASIEWICZ JAN, *Du principe de contradiction chez Aristote*, trad. fr. de Dorota Sikora, préface de Roger Pouivet, Paris, L'Éclat, 2000 (œuvre originale publiée en 1910).
- [123] LYCAN WILLIAM G., *Consciousness*, First published October 1, 1987.
- [124] MAGNIN THIERRY, *La Mesure humaine dans le post et le transhumanisme*, dans Revue d'éthique et de théologie morale, 2019-3 (Num. 303), pages 53 à 68.
- [125] MAGNIN THIERRY, *De la robustesse du cyborg à la vulnérabilité de l'être humain*, Enseignement catholique, <https://enseignement-catholique.fr/wp-content/uploads/2018/05/MAGNIN-ethique-cyborg-CCIC.pdf> (consulté le 5 novembre 2014).

- [126] MARCHETTI GIORGIO, *Consciousness : a unique way of processing information*, Cognitive Processing (2018) 19 :435–464.
- [127] MANSION SUZANNE, *La Critique de la Théorie des Idées dans le ΠΕΡΙ ΙΔΕΩΝ*, Revue philosophique de Louvain, 1949.
- [128] MCCARTHY J., MINSKY M. L., ROCHESTER N., SHANNON C.E. (1955), *A Proposal for the Dartmouth Summer Research Project On Artificial Intelligence*, Bell Telephone Laboratories, 31 Août 1955, <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html> (consulté le 17 octobre 2024).
- [129] MCCULLOCH, W.S., PITTS, W., *A Logical Calculus of the Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics 5, 115–133, 1943.
- [130] MCGINN COLIN, *Can We Solve the Mind-Body problem ?*, 1989, Mind, New Series, Volume 98, Issue 391 (Jul., 1989).
- [131] MERLEAU-PONTY MAURICE, *Phénoménologie de la perception*, Paris, Gallimard, coll. « Tel », 2005.
- [132] MONIER CYRIL, *Les neurosciences au sein des sciences de la cognition : Une revue de la littérature plaidant pour un devenir spinoziste des neurosciences*, 2019.
- [133] MUNTEAN IOAN & HOWARD DON, *Artificial Moral Agents : Creative, Autonomous, Social. An Approach Based on Evolutionary Computation*, in JOHANNA SEIBT, RAUL HAKLI & MARCO NORSKOV, *Sociable Robots and the Future of Social Relations : Proceedings of Robo-Philosophy*, IOS Press, 2014.
- [134] NABOKOV, *Lolita*, annotée par ALFRED APPEL JR., New York, Vintage Books, 1991.
- [135] NAGEL THOMAS, *What Is It Like to Be a Bat ?*, The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.
- [136] A. NEWELL, J. C. SHAW, H. A.SIMON, *Report on a General Problem-Solving rogram*, 30 décembre 1958, révisé le 9 février 1959.
- [137] O’NEIL CATHY, *Algorithmes : la bombe à retardement*, Broché – Grand livre, 7 novembre 2018
- [138] OPILLARD THIERRY, *La Dégradation Cambridge : Même pas vrai ?*, Les Actes de Lecture n.102, juin 2008.
- [139] ORTIGUES EDMOND, *Edelman et la conscience humaine*, Éditions de Minuit, dans *Critique* 2005-12 (Num. 703), 2005-12 (Num. 703), pages 917 à 926.
- [140] PAPADIMITRIOU CHRISTOS H., LEWIS HARRY. R., *Elements of the Theory of Calculation*, Second Edition, Prentice-Hall, Upper Saddle River, New Jersey, 1998.

- [141] PARMÉNIDE, *Fragment du poème*, <http://philoctetes.free.fr/uniparmenide.htm> (consulté le 10 novembre 2024).
- [142] PASCAL BLAISE, *Pensées*, L'édition électronique de DESCOTES D. ET PROUST G., <https://www.penseesdepascal.fr/> (consulté le 17 octobre 2014).
- [143] PATY MICHEL, *Le vide matériel, ou la matière crée l'espace*, Simon Diner et Edgard Gunzig, Univers du tout et du rien, Éditions de l'Université Libre de Bruxelles/Complexe, p. 22-44, 1998.)
- [144] PEÑA ADOLFO, *The Dreyfus Model of Clinical Problem-Solving Skills Acquisition : a Critical Perspective*. Med Educ Online, juin 2010.
- [145] PENFIELD W., BOLDREY E, *Somatic Motor and Sensory Representation in the Cerebral Cortex of Man as Studied by Electrical Stimulation*, Brain, 1937 et PENFIELD W, RASMUSSEN T., *The cerebral cortex of man*, New York : Macmillan, 1950.
- [146] GEORGES PEREC, *La vie, mode d'emploi*, Le Livre de Poche, 1er janvier 1980.
- [147] *The Perl Design Patterns Book*, GNU FDL license, 1995.
- [148] PERRIN MICHEL, *Logique « sauvage » des psychotropes : le cas des sociétés chamaniques*, Psychotropes, 1991, VI, 4, pp. 115-123.
- [149] PERRIN DANIEL, *Les Géométries non euclidiennes*, Conférence de Romilly, <https://www.imo.universite-paris-saclay.fr/~daniel.perrin/Conferences/Romilly.pdf> (consulté le 17 octobre 2014).
- [150] PIAGET JEAN, *Language Within Cognition : Schemes of Action and Language Learning*. In M. Piatelli-Palmerini (Ed.), *Language and learning*. (pp. 163–183). Cambridge, MA : Harvard University Press, 1980.
- [151] PLATON, *Ménon*, trad. Cousin, Philosophie, Septembre 2009.
- [152] PLATON *République*, trad. de COUSIN, consultable sur *L'Antiquité grecque et latine du moyen âge*, <https://remacle.org/> (consulté le 3 novembre 2024).
- [153] PLATON, *Théétète [ou sur la science ; genre peirastique]*, trad. CHAMBRY E., La Bibliothèque électronique du Québec, Collection Philosophie, Volume 9 : version 1.01.
- [154] PLUTARQUE, *Dialogues pythiques*, trad. ILDEFONSE FRÉDÉRIQUE, Éditions Flammarion, Paris, 2006.
- [155] POINCARÉ HENRY, *Les Mathématiques et la Logique*, Revue de Méta-physique et de morale, années 1905, pp. 815-835, 1906, pp. 17-38 et pp. 297- 317.
- [156] POINCARÉ HENRY, *La Science et l'Hypothèse*, Flammarion, Collection Bibliothèque de philosophie scientifique, Paris, 1917.

- [157] POIRIER P., PAYETTE N., *Les Gardiens du bon usage*, Philosophiques, vol.34, n.1, pp. 183-200.
- [158] POTGIETER, PETRUS H., *Zeno machines and hypercomputation*, in Theoretical Computer Science, Volume 358, Issue 1, 31 July 2006, Pages 23-33
- [159] PREMACK DAVID, WOODRUFF GUY, *Does the Chimpanzee Have a Theory of Mind ?*, Behavioral and Brain Sciences, Vol. 1, N. 4, décembre 1978, pp. 515-526.
- [160] PRINZ JESSE J., *Level-Headed Mysterianism and Artificial Experience*, Journal of Consciousness Studies 10 (4-5) :111-132, 2003.
- [161] PROCLUS, *Commentaires sur le premier livre des Éléments d'Euclide*, trad. P. Ver Eecke, Bruges, 1948, rééd. 1959. Influencé par Jamblique (Collection des dogmes pythagoriciens, III : Sur la science commune des mathématiques).
- [162] PTOLÉMÉE CLAUDE, *Almageste*.
- [163] RUDIN CYNTHIA, WANG CAROLINE, COKER BEAU, *The age of secrecy and unfairness in recidivism prediction*, 2018, arXiv :1811.00731.
- [164] QUINCEY THOMAS DE, *Les derniers jours d'Emmanuel Kant*, éditions Mille et une nuits, 1996.
- [165] QUINE W. V. O., *Word and Object*, Cambridge, MA : MIT Press, 1960.
- [166] RAMUS FRANK, *Un épouvantail bien commode*, Presses Universitaires de France, 2014.
- [167] REILLY DOUGLAS L., COOPER LEON N., ELBAUM CHARLES, *A Neural Model for Category Learning*, Biological Cybernetics, 45, 35-41, 1982.
- [168] RISSER ANTHONY, CLAVERT PHILIPPE, MIGAUD HENRI et al., *Quel est le taux de texte généré par de l'intelligence artificielle sur une année de publication dans Orthopaedics & Traumatology : Surgery & Research ? Analyse de 425 articles avant vs après la mise en ligne de ChatGPT en novembre 2022*, Revue de chirurgie orthopédique et traumatologique, 109(3).
- [169] ROELOFS COZ, ZEEMAN WPC, *Influence of Different Sequences of Optical Stimuli on the Estimation of Duration of a Given Interval of Time*. Acta Psychologica, 1951 ;8 :89-128.
- [170] ROSE, D., SUMMERS, J. *Duration Illusions in a Train of Visual Stimuli*, Scopus, Volume 24, Issue 10, 1995, Pages 1177-1187.
- [171] ROSENFELD BORIS, *A History of Non-Euclidean Geometry : Evolution of the Concept of a Geometric Space*, Springer, New York, 1988.
- [172] ROUDAUT FRÉDÉRIC, *Comment on invente les hypothèses : Peirce et la théorie de l'abduction*, dans Cahiers philosophiques 2017-3 (Num 150), pages 45 à 65.

- [173] ROUX SYLVAIN, *La recherche du Principe chez Platon, Aristote et Plotin*, Vrin, 2004, <https://books.openedition.org/vrin/5432?lang=fr> (consulté le 17 octobre 2024).
- [174] RUMELHART DAVID E., MCCLELLAND JAMES L., *Parallel Distributed Processing. Exploration in the microstructure of cognition, vol I*. Cambridge, MIT Press.
- [175] RUSSELL BERTRAND, *Set Theory as Based on the Theory of Types* (1910, 1912, 1913) *Principia Mathematica*, 3 volumes, Cambridge : Cambridge University Press ; 2nd edition, 1925 (Vol. I), 1927 (Vols II, III).
- [176] RUSSELL BERTRAND ET WHITEHEAD ALFRED NORTH (1910, 1912, 1913) *Principia Mathematica*, 3 volumes, Cambridge : Cambridge University Press ; 2nd edition, 1925 (Vol. I), 1927 (Vols II, III).
- [177] RUSSELL STUART ET NORVIG PETER, *Artificial Intelligence : A Modern Approach*, Pearson Education, troisième édition, 2003.
- [178] RYLE G., *The Concept of Mind*, Hutchinson, London, 1949.
- [179] SAERTENER OLIVIER, GUAY ALEXANDRE, *A New Look at Surveillance. Or when After is Different*, European Journal of Philosophy of Science, 2016.
- [180] ROBERTSON NEIL, SANDERS DANIEL P., SEYMOUR PAUL et THOMAS ROBIN, *A New Proof of the Four-Colour Theorem*, Electronic Research Announcements of the American Mathematical Society, Volume 2, Number 1, August 1996.
- [181] SARTRE, J-P., *La Transcendance de l'Ego : Esquisse d'une description phénoménologique*, Librairie Philosophique J.Vrin, Paris, 1966.
- [182] WALTER SAVITCH, *Relationships Between Nondeterministic and Deterministic Complexities*, Journal of Computer and System Sciences, vol. 4, no. 2, 1970.
- SÉGUI JEAN, *Toujours la même histoire*, Éditions Non Standard, 2017.
- [183] SCHIFFMAN H. R., BOBKO D. J., *Effects of Stimulus Complexity on the Perception of Brief Temporal Intervals*, Journal of Experimental Psychology, 1974;103 :156–159.
- [184] Schopenhauer A., *Le monde comme volonté et représentation* (Die Welt als Wille und Vorstellung, 1813), traduit par Auguste Burdeau, édition revue par Richard Roos, Préface de Clément Rosset, Paris, PUF / Quadrige, 2014.
- [185] SIMON HERBERT, *The New Science of Management Decision*, édition révisée, Englewood Cliffs, NJ : Prentice-Hall, 1977.
- [186] SOUCHON MARGUERITE, *Le Dieu de Dostoïevski*, Éditions Première partie, 6 octobre 2021.

- [187] STANISLAW LEŚNIEWSKI, *Über die Grundlagen der Ontologie* [On the Foundations of Ontology]. Sprawozdania z posiedzeń Towarzystwa Naukowego Warszawskiego, Wydział III [Comptes rendus des séances de la Société des Sciences et des Lettres de Varsovie, Classe III], 22 (1930), 111–132.
- [188] TAYLOR, *Paying attention to consciousness*, Progress in Neurobiology, 2008.
- [189] THOMASSON DAN, *Georges Perec's 10x10 Knight's Tour*, consulté le 17 octobre 2024, <https://www.danthomasson.com/georges-perec.html>.
- [190] TILLEMANS T., *Dharmakirti* in Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*, Spring 2021 ed. Metaphysics Research Lab, Stanford University ; 2021.
- [191] TONONI G., *Integrated Information Theory of Consciousness : an Updated Account*, Arch Ital Biol 150 : 56-90, 2012.
TONONI GIULIO, HASHMI A, NERE A., *Sleep-Dependent Synaptic Down-Selection (II) : Single-Neuron Level Benefits for Matching, Selectivity, and Specificity*, Front Neurol 4 :148, 2013.
- [192] TONONI GIULIO, BOLY MELANIE, MASSIMINI MARCELLO, KOCH CHRISTOF, *Integrated Information Theory : From Consciousness to its Physical Substrate*, Nature Reviews Neurosciences, Mai 2016.
- [193] TONONI GIULIO et al., *Integrated Information Theory (IIT) 4.0 : Formulating the Properties of Phenomenal Existence in Physical Terms*, arXiv, 30 décembre 2022.
- [194] TONONI GIULIO et al., *Only What Exists Can Cause : An Intrinsic View on Free Will*, arXiv, Juin 2022.
- [195] TONONI GIULIO, *On being*, à paraître.
- [196] TURING ALAN (1950) *Computing Machinery and Intelligence*, Mind 49 : 433-460.
- [197] TURING ALAN, *Intelligent Machinery*, Collected Works of A.M. Turing, Mechanical Intelligence, 1948.
- [198] ULLMAN, SHIMON, *Tacit Assumptions in the Computational Study of Vision* in A. Gorea (ed.), *Representations of Vision : Trends and Tacit Assumptions in Vision Research*, Cambridge, England : Cambridge University Press, 305-317, 1991. Hungarian translation, 1996.
- [199] VAN DER LINDEN TIM, *Notions de logique mathématique*, note de cours de l'UCLouvain, 19 mai 2013.
- [200] VAN GULICK R., *What would count as explaining consciousness ?* In T. Metzinger, ed. *Conscious Experience*. Paderborn : Ferdinand Schöningh, 1995.
- [201] VARELA FRANCISCO, *La Recherche*, No. 308 Avril 1998.

- [202] VERLEYSSEN MICHEL, FRANÇOIS BLAYOT, *Les réseaux de neurones artificiels (Que sais-je ?)*, Presses Universitaires de France, 1996.
- [203] VELMANS MAX, *Understanding Consciousness*, second edition, Routledge/Psychology Press, 2009.
- [204] VINCENT L., *Comment devient-on amoureux ?*, Paris, Odile Jacob, 2004
- [205] WERNICKE C., *Grundriss der Psychiatrie in klinischen Vorlesungen*, 1900.
- [206] WIENER NORBERT, *Cybernetics : Or Control and Communication in the Animal and the Machine*, Preface, page xi (footnote 1), 1948.
- [207] WITTGENSTEIN LUDWIG, *Tractatus Logico-philosophicus*, Wilhelm Ostwald's Annalen der Naturphilosophie, Vienne, 1921, traduit par OGDEN C.K., *Frank Ramsey* et Wittgenstein lui-même en 1922.
- [208] YEVAROUSKY VALERY, *The "Kantian Brain" Concept : Modeling the Limit of Human Perception and Existence*, septembre 2023.
- [209] ZERMELO, *Investigations in the Foundations of Set Theory I* (Abridged Version), Trad. Andreas Reif, 1908.

Annexes

Annexe 1 : Une nouvelle écrite par ChatGPT

Les progrès de ChatGPT ont, à juste titre sans doute, défrayé la chronique. Depuis sa sortie en 2022, l'IA n'a pas cessé de s'améliorer et un nombre toujours croissant de pages web sont rédigées avec ChatGPT¹⁹. À ce titre, le phénomène représente peut-être une menace existentielle inaugurale à la suprématie et au bon fonctionnement des moteurs de recherche traditionnels. Afin d'illustrer quelques-unes de ces forces et des limitations de ChatGPT, le lecteur trouvera ci-dessous une nouvelle intégralement générée par ce logiciel (version o1 preview) à la date du 17 septembre 2024, précédée des consignes ayant été introduites dans le programme.

Cette expérience a été menée en coopération avec Joseph Lange, cofondateur de Fabuly, une start-up spécialisée dans la création de livres audio qui organisait un concours de nouvelles en septembre 2024. ChatGPT s'est vu introduire les consignes de ce concours²⁰. Le résultat a ensuite été traduit en français en utilisant ChatPGT lui-même. La nouvelle porte pour titre « La Tisserande des murmures ». On remarquera certes des naïvetés évidentes en ce qui concerne le contenu, particulièrement quant à son caractère émotionnel, mais les pires détracteurs de l'intelligence artificielle ne sauront nier la présence d'une structure narrative claire.

19. Le phénomène, du reste, ne touche pas que les pages web : une étude publiée par Elsevier démontre que, en moyenne, 5,6% du contenu des articles recensés en novembre 2023 sur *Orthopaedics & Traumatology : Surgery & Research (OTSR)*, une plateforme de publications scientifiques à vocation médicale, avaient été générés par ChatGPT. Cet usage concerne en grande partie les « abstracts » (cf. *Quel est le taux de texte généré par de l'intelligence artificielle sur une année de publication dans Orthopaedics & Traumatology : Surgery & Research ?*)

20. J'ai moi-même rédigé les consignes du concours, en échange de quoi j'ai demandé à conduire quelques expériences avec la version payante de ChatGPT utilisée par Fabuly.

Consignes de la nouvelle

To Your Pens !

Whether you're an aspiring writer or an established author, an aesthete, a poet, a novelist, or a self-repressed bookworm, we invite you to reveal your hidden talents in Fabuly's inaugural writing contest.

With the Christmas holidays fast approaching, you'll have two months to craft a 2,000-word story on the theme of "an unexpected encounter." The top ten entries will be published on our platform and shared with thousands of readers around the globe.

The grand prize winner, chosen by our panel of judges, will receive 500 USD. Their story will not only be highlighted but also professionally illustrated and produced in an audio format. All winners will have the opportunity to include a brief biography and a link to their website or preferred social media profile alongside their story.

A Web of Unlikely Encounters

Human paths often resemble absurd canvases strewn with choices, faces and disjointed words, which cruel Muses surreptitiously blend together. An encounter that seems insignificant at a given moment can turn out to be either a stray bullet or a time bomb : it traverses time and space, passing from hand to hand, before exploding in one's face. The truth becomes clear : this encounter was meant to happen ; it was a matter of inscrutable necessity.

The art of extracting meaning from the most far-fetched encounters, the relentless squeezing, wrenching, grasping and covetous irony of life, might be what inspires Goethe to say :

"Nobody must and can unveil the mysteries of human existence ; there are stones on the path of life against which every traveler stumbles. It is up to the poet to point them out." (Maxims and Reflections)

In this writing contest, we ask you to vividly depict one of these stones for the reader. May this stone impede the traveler, sometimes against their will, in such a way as to subtly or deeply alter their journey ! The unexpected encounter can occur in a fantastic or realistic setting. It may emerge in the form of a human, an animal, or a magical creature. A betrayed lover, a hopeless romantic, a forgotten parent, a deceased (wealthy) friend, or a fleeting glance exchanged with a stranger — numerous triggers can thrust the hero into a wholly different reality.

To inspire you, we offer a few quotes from great literary masterpieces. You will find a wide variety of situations, settings, characters, and impacts on the narrative. You can explore the original works of these quotes on Fa-

bully platform :

Quote 1 :

I was walking along, my eyes on the ground. Suddenly I heard a voice ; I looked across the fence, and was thunder-struck. . . . I was confronted with a curious spectacle. A few paces from me on the grass between the green raspberry bushes stood a tall slender girl in a striped pink dress, with a white kerchief on her head ; four young men were close round her, and she was slapping them by turns on the forehead with those small grey flowers, the name of which I don't know, though they are well known to children ; the flowers form little bags, and burst open with a pop when you strike them against anything hard. (First Love, Turgenev)

Quote 2 :

But why, he was always asking himself, why had such an important, such a decisive and at the same time such an absolutely chance meeting happened in the Hay Market (where he had moreover no reason to go) at the very hour, the very minute of his life when he was just in the very mood and in the very circumstances in which that meeting was able to exert the gravest and most decisive influence on his whole destiny ? As though it had been lying in wait for him on purpose ! (Crime and Punishment, Dostoyevsky)

Quote 3 :

The cellar-door flew open with a booming sound, and then he heard the noise much louder, on the floors below ; then coming up the stairs ; then coming straight towards his door. "It's humbug still !" said Scrooge. "I won't believe it." His colour changed though, when, without a pause, it came on through the heavy door, and passed into the room before his eyes. Upon its coming in, the dying flame leaped up, as though it cried, "I know him ; Marley's Ghost !" (A Christmas Carol, Charles Dickens)

Quote 4 :

I think everyone expected to see a man emerge — possibly something a little unlike us terrestrial men, but in all essentials a man. I know I did. But, looking, I presently saw something stirring within the shadow : greyish billowy movements, one above another, and then two luminous disks— like eyes (The War of

the Worlds, H. G. Wells)

Quote 5 :

What would have happened if she had not lost that necklace? Who knows? who knows? How life is strange and changeful! How little a thing is needed for us to be lost or to be saved! But, one Sunday, having gone to take a walk in the Champs Élysées to refresh herself from the labors of the week, she suddenly perceived a woman who was leading a child. (The Necklace, Maupassant)

Quote 6 :

One night as I sat, half stupefied, in a den of more than infamy, my attention was suddenly drawn to some black object, reposing upon the head of one of the immense hogsheads of gin, or of rum, which constituted the chief furniture of the apartment. I had been looking steadily at the top of this hogshead for some minutes, and what now caused me surprise was the fact that I had not sooner perceived the object thereupon. I approached it, and touched it with my hand. It was a black cat—a very large one. (The Black Cat, Edgar Allan Poe)

Quote 7 :

The King turned round, and saw a bearded man come running out of the wood. The man held his hands pressed against his stomach, and blood was flowing from under them. When he reached the King, he fell fainting on the ground moaning feebly. (Three questions, Tolstoy)

La Tisserande des murmures

La ville bourdonnait de son rythme habituel—klaxons retentissants, sirènes lointaines, et le murmure de conversations innombrables se fondant en un seul bruit indistinct. Amelia naviguait à travers la foule sur le trottoir, ses pensées consumées par les routines monotones qui avaient fini par définir son existence. Restauratrice de textiles anciens au musée, son travail convenait à sa nature méticuleuse mais laissait en elle un désir d'autre chose.

Un soir, en cataloguant une collection d'objets donnés, elle découvrit une pièce singulière—un fragment de tapisserie antique, ses bords effilochés et ses couleurs fanées. Le design était complexe : un labyrinthe de fils entrelacés formant des motifs qui semblaient changer sous son regard. Il n'y avait

aucune trace de son origine, aucune note sur le donateur. Intriguée, Amelia décida de l'examiner de plus près.

Sous la douce lueur de sa lampe de bureau, elle remarqua un symbole à peine visible tissé dans un coin—un petit sablier stylisé enlacé de lierre. Cela raviva un souvenir d'enfance, une histoire que sa grand-mère lui racontait sur une légendaire tisserande dont les créations étaient censées posséder des pouvoirs mystiques. Rangeant cela au rang des idées fantaisistes, elle mit la tapisserie de côté.

Cette nuit-là, le sommeil la fuyait. Des images de la tapisserie dansaient derrière ses paupières closes, l'appelant à revenir. Cédant à cette attraction, elle s'habilla et retourna au musée. Les rues de la ville étaient étrangement silencieuses, un contraste frappant avec leur frénésie diurne. Le gardien du musée, habitué à ses visites nocturnes, lui fit un signe complice alors qu'elle signalait le registre.

De retour dans l'atelier, Amelia déploya la tapisserie sur la table. En passant ses doigts sur le tissu, une sensation de picotement parcourut son corps. L'air s'épaissit légèrement, et la pièce sembla émettre un léger bourdonnement, bien que la source en soit indiscernable. Surprise, elle recula.

« Qui est là ? » appela-t-elle d'une voix à peine audible.

Seul le silence lui répondit. Chassant son malaise, elle se pencha à nouveau pour inspecter la tapisserie. Cette fois, les motifs semblaient différents—des changements subtils dans le tissage qu'elle ne pouvait tout à fait saisir. Déterminée à documenter ses observations, elle prit son carnet de croquis.

Les heures s'écoulèrent tandis qu'elle dessinait méticuleusement chaque détail. Lorsqu'elle leva enfin les yeux, l'aube se levait, la lumière filtrant à travers les stores. Épuisée mais satisfaite, elle rassembla ses affaires pour partir. Alors qu'elle éteignait les lumières, une lueur attira son attention. La tapisserie semblait émettre une faible luminescence, mais en regardant de plus près, elle redevint ordinaire.

Au cours des jours suivants, Amelia se retrouva préoccupée par la tapisserie. Elle commença à rechercher le symbole du sablier, plongeant dans les archives et les textes obscurs. Un après-midi, nichée dans un coin poussiéreux de la bibliothèque, elle tomba sur une référence dans un vieux tome sur les artisans oubliés.

« La Tisserande des Murmures », lut-elle à voix haute, « une figure mythique dont les tapisseries étaient censées relier les royaumes, connectant le passé, le présent et le futur. » L'illustration associée portait le même sablier enlacé de lierre. Un frisson parcourut son échine.

Ce soir-là, alors qu'elle réfléchissait aux implications, son téléphone vibra avec un message de son ami Leo, un historien passionné par l'ésotérisme.

« J'ai trouvé quelque chose qui pourrait t'intéresser », disait le message, suivi d'une adresse.

Sa curiosité piquée, elle accepta de le rejoindre. L'adresse la mena à une

boutique d'antiquités qu'elle n'avait jamais remarquée auparavant, nichée entre une boulangerie et une librairie. La cloche au-dessus de la porte tinta doucement lorsqu'elle entra.

« Amelia, par ici », l'appela Leo depuis le fond du magasin. Il se tenait à côté d'une série de textiles accrochés au mur. « Ça te dit quelque chose ? »

Ses yeux s'élargirent en reconnaissant les motifs—similaires à ceux de la tapisserie qu'elle étudiait. « Où as-tu trouvé cela ? »

« La propriétaire les a hérités d'un parent. Je me suis dit que tu voudrais y jeter un œil. »

Une femme âgée émergea de derrière un rideau. Ses yeux étaient d'un vert saisissant, pétillant d'une énergie qui démentait son âge.

« Vous devez être Amelia », dit-elle chaleureusement. « J'ai entendu beaucoup de bien de votre travail. »

Surprise, Amelia répondit : « Je ne savais pas que ma réputation s'étendait au-delà du musée. »

La femme sourit énigmatiquement. « Certaines choses sont destinées à être découvertes. »

Amelia examina les textiles, notant les motifs récurrents. « Ces designs—ils sont liés à une légende que j'étudie. »

« Ah, la Tisserande des Murmures », acquiesça la femme. « Une histoire aussi ancienne que le temps lui-même. »

« En savez-vous davantage ? » demanda Amelia avec enthousiasme.

« Peut-être. Mais la connaissance se révèle souvent à ceux qui sont prêts à la recevoir. »

Amelia hésita. « Que voulez-vous dire ? »

La femme désigna la tapisserie. « Ces fils contiennent plus que du simple tissu. Ils sont un conduit, un lien vers quelque chose de plus grand. »

Leo rit doucement. « Elle est toujours aussi mystérieuse. »

La femme lui lança un regard malicieux. « Patience, Leo. Maintenant, Amelia, accepteriez-vous de partager une tasse de thé ? Nous avons beaucoup à discuter. »

Ils s'assirent à une petite table près de la fenêtre tandis que la femme, qui se présenta comme Seraphine, versa du thé parfumé dans de délicates tasses.

« La tapisserie que vous avez trouvée fait partie d'une collection créée par une artisane bien en avance sur son temps », commença Seraphine. « Elle tissait des pièces qui connectaient différents points de l'existence, des fils capables d'influencer la trame même de la réalité. »

Amelia écoutait attentivement, le scepticisme et la fascination se disputant en elle. « Vous suggérez que la tapisserie a... des capacités ? »

« Des méthodes avancées peuvent sembler magiques à ceux qui ne les connaissent pas », répondit Seraphine avec douceur. « Dites-moi, avez-vous vécu des choses inhabituelles depuis que vous l'avez trouvée ? »

Elle raconta les événements étranges—le bourdonnement dans la pièce fermée, les motifs changeants, la faible lueur. Seraphine acquiesça en connaissance de cause.

« La tapisserie réagit à vous », dit-elle. « Elle sent votre curiosité, votre ouverture aux possibilités. »

Amelia ressentit un mélange d'excitation et d'appréhension. « Que devrais-je faire ? »

« Acceptez-la », encouragea Seraphine. « Laissez-la vous guider. »

Cette nuit-là, de retour au musée, Amelia se tint devant la tapisserie. Prenant une profonde inspiration, elle tendit la main et toucha le symbole du sablier. La pièce sembla onduler, et soudain, elle ne se trouvait plus dans l'atelier.

Elle se retrouva dans une vaste étendue sous un ciel étoilé. L'air était vif, rempli de senteurs inconnues mais réconfortantes. Au loin, une silhouette approchait—une femme vêtue de vêtements ornés des mêmes motifs complexes.

« Bienvenue, Amelia », dit la femme, sa voix résonnant doucement. « J'attendais ce moment. »

« Qui êtes-vous ? » demanda Amelia, le cœur battant.

« Je suis Elara », répondit-elle. « Gardienne des fils. »

Amelia sentit une vague de calme l'envahir. « Pourquoi suis-je ici ? »

« Parce que vous avez le potentiel pour poursuivre mon œuvre », dit Elara. « Les connexions qui lient la réalité ont besoin d'attention. »

« Je ne comprends pas », dit Amelia en secouant la tête.

« Vous comprendrez, avec le temps », assura Elara. « Chaque action, chaque choix, tisse la trame de ce qui est et de ce qui pourrait être. »

Avant qu'elle ne puisse en demander davantage, la scène se dissipa doucement, et Amelia était de retour dans l'atelier. La tapisserie était devant elle, ses couleurs désormais plus vives, les motifs se mouvant subtilement comme s'ils étaient vivants.

Submergée, elle s'assit, essayant de comprendre ce qui venait de se passer. La partie rationnelle de son esprit luttait avec l'improbabilité de l'expérience, mais au fond d'elle, elle ressentait une nouvelle vocation.

Dans les jours qui suivirent, Amelia commença à expérimenter, tissant de petites pièces inspirées de la tapisserie. Elle remarqua des changements subtils—un collègue retrouvant l'inspiration après des semaines de stagnation, une amie vivant une opportunité inattendue. Les événements semblaient mineurs mais significatifs, suggérant une influence plus grande.

Un après-midi, Leo lui rendit visite au musée. « Tu sembles différente », remarqua-t-il.

« Vraiment ? » répondit-elle distraitement, ses doigts manipulant habilement les fils.

« Il y a une étincelle en toi », dit-il. « Comme si tu avais découvert un secret. »

Elle sourit doucement. « Peut-être bien. »

Il jeta un coup d'œil à la tapisserie. « Tu continues d'explorer ses mystères ? »

« D'une certaine manière », dit-elle. « Mais il s'agit plutôt de créer de nouveaux motifs maintenant. »

Il acquiesça pensivement. « Je voulais te remercier. L'idée que tu m'as donnée a débloqué une partie importante de mes recherches. »

Elle leva les yeux, surprise. « Je ne t'ai pas donné d'idée. »

Il inclina la tête. « Tu m'as envoyé cette référence sur les systèmes interconnectés. »

La réalisation l'envahit. Le petit tissage qu'elle avait laissé sur son bureau, orné de motifs complexes—il avait d'une manière ou d'une autre transmis des informations.

« De rien », dit-elle simplement.

Au fil du temps, Amelia embrassa son nouveau rôle. Elle continua son travail au musée, mais passait ses soirées à tisser des fils qui influençaient subtilement le monde autour d'elle. Elle restait en contact avec Seraphine, qui la guidait tout en l'encourageant à explorer par elle-même.

Une nuit, alors qu'elle rentrait chez elle, elle remarqua un homme debout seul, le regard perdu et troublé. Instinctivement, elle plongea la main dans son sac et sortit un petit charme tissé.

« Excusez-moi », dit-elle doucement. « Je pense que ceci pourrait vous être utile. »

Il la regarda, la surprise dans les yeux. « Pourquoi me le donnez-vous ? »

Elle sourit. « Parfois, un petit geste peut nous rappeler que nous ne sommes pas seuls. »

Il accepta le charme avec hésitation. En le tenant, un changement subtil s'opéra en lui—la tension dans ses épaules s'atténua, une lueur de clarté apparut dans son regard. « Merci », murmura-t-il.

Elle acquiesça et poursuivit son chemin, se fondant dans le flot des passants. Derrière elle, l'homme se redressa légèrement, une résolution nouvelle visible dans sa posture.

Amelia réalisa que son voyage s'était entremêlé avec celui des autres, chaque interaction étant un ajustement subtil dans la complexe tapisserie de l'existence. Elle ne se sentait plus perdue dans la monotonie du quotidien ; au contraire, elle faisait partie de quelque chose de profond, collaborant silencieusement avec des forces qu'elle commençait tout juste à comprendre.

Un matin, elle reçut un colis au musée—un livre sur des techniques de tissage avancées, rempli de motifs et de symboles à la fois anciens et futuristes. À l'intérieur, une note : « Pour la Tisserande des Nouveaux Chemins. Continue d'explorer.—S. »

Amelia sourit, glissant le livre sous son bras. Alors qu'elle se déplaçait dans les halls familiers, elle ressentit une profonde connexion avec tout ce qui l'entourait—les artefacts, les personnes, l'essence même du monde. La

tapisserie qui était autrefois une énigme était désormais une passerelle vers d'innombrables possibilités.

Elle sortit dans la ville animée, la cacophonie des sons formant désormais une symphonie harmonieuse. Chaque visage qu'elle croisait portait des histoires inexprimées, chaque moment était un point de bascule potentiel. Le monde regorgeait de motifs attendant d'être compris, de fils attendant d'être tissés.

Amelia prit une profonde inspiration, sentant le pouls de la vie autour d'elle. Avec un sens renouvelé du but, elle s'avança, prête à embrasser le dessein complexe de ce qui l'attendait.

Annexe 2 : Hypothèses détaillées du Théorème de Gödel

Les conditions exactes des théorèmes de Gödel, quoique peu exigeantes, sont rarement mentionnées sous leur forme la plus complète et la plus rigoureuse. J'ai pris la peine de les esquisser ici, sous une forme certes non technique, mais simple et concise. Le lecteur pourra les trouver sous une forme encore plus exhaustive dans *Les Limitations internes des formalismes*²¹ :

1. Le système formel est exprimé dans un langage du premier ordre.
2. Il est cohérent.
3. Sa syntaxe peut s'exprimer au moyen de l'arithmétique récursive.
4. Il est au moins aussi puissant que le système PA.

La condition 3 revient à dire qu'il est possible de dresser une correspondance biunivoque entre les éléments du système et les nombres entiers. Chacune des constantes, des variables, des fonctions et des prédicats doit se voir attribuer un entier distinct. De plus, les propositions doivent correspondre à un entier qui soit fonction univoque des entiers associés aux éléments dont elles sont constituées. Un exemple de telle correspondance est le codage de Gödel.

Le système doit en outre satisfaire les conditions suivantes : a) la propriété pour une proposition d'être un axiome doit être représentable récursivement ; et b) la propriété, pour une proposition, d'être la conséquence immédiate d'une ou plusieurs autres propositions doit être représentable récursivement.

La condition 4 peut être explicitée à son tour au moyen de quatre clauses :

- Le système doit comporter un opérateur de négation $\neg$ muni de son interprétation usuelle.

21. LADRIERE, pp. 160-163.

- Le système doit comporter le quantificateur universel $\forall$ muni de son interprétation usuelle.
- Le système doit comporter un opérateur d'implication $\rightarrow$ muni de son interprétation usuelle. Il doit en outre être doté d'une règle d'inférence qui précise le rôle de l'implication par rapport à la notion de dérivabilité :

$$\frac{A \quad (A \rightarrow B)}{B}$$

- Le système doit comporter une formalisation de l'arithmétique récursive, ce qui signifie :
 - a) qu'il comporte des expressions qui représentent sans ambiguïté les nombres entiers ;
 - b) qu'il comporte des symboles de variables ainsi que des symboles de fonctions et de prédicats ayant ces variables pour arguments ;
 - c) qu'il comporte un opérateur de substitution qui permet de remplacer, au sein d'une fonction ou d'un prédicat, ces variables par des expressions correspondant à des entiers ;
 - d) que tous les prédicats récursifs de l'arithmétique récursive sont représentables dans le système formel. Autrement dit, tout prédicat récursif $P(x_1, \dots, x_n)$ doit correspondre à une expression prédictive $A(N_1, \dots, N_n)$, où $N_1, \dots, N_n$ sont les expressions du système représentant les entiers $x_1, \dots, x_n$.

Annexe 3 : La Thèse de Church-Turing

Pour de plus amples détails concernant ce sujet, le lecteur est renvoyé au chapitre 6 des *Limitations internes des formalismes*^[22], dont cette annexe s'inspire.

Méthode effective

Peu après la parution des théorèmes d'incomplétude s'est posé la question de savoir s'il existait une « méthode effective » pour déterminer la démontrabilité d'une proposition quelconque. Cette interrogation est ce qui définit la *résolubilité* d'un système formel.

Définition 29. *Une théorie est dite résoluble (au sens syntaxique) s'il existe une "méthode effective" permettant de décider, pour toute proposition ϕ du langage, si ϕ est dérivable ou non. Le problème qui consiste à trouver une telle méthode est appelé problème de la décision*^[23].

Ce qui doit être entendu par "fonction effective" peut être interprété de différentes manières. Pas moins de trois notions ont été avancées pour capturer cette intuition^[24]. La Thèse de Church-Turing stipule que ces approches sont équivalentes et qu'elles capturent effectivement le concept de « calculabilité ». Autrement dit, tout ce qui est calculable peut s'exprimer par l'une de ses trois voies, et réciproquement.

1. Fonction Turing-calculable
2. Fonction récursive générale
3. Fonction λ -définissable

Fonction Turing-calculable Une machine de Turing comporte un ruban, un organe de lecture, un organe d'exécution et une mémoire. Le ruban,

22. LADRIÈRE , pp. 227-278.

23. LADRIÈRE, *Ibidem*, p. 61.

24. LADRIÈRE, *Ibidem*, p. 231-235.

supposé infiniment long, est divisé en cellules. Chacune d'entre elles contient un symbole parmi un alphabet fini, ou un "0" signifiant que la cellule est vide. L'organe de lecture déchiffre le symbole de la cellule sur lequel il se trouve. Il peut se déplacer sur le ruban vers la gauche ou vers la droite, en fonction des instructions qui lui sont données par l'organe d'exécution. Il peut également effacer le signe de la cellule, ou en imprimer un nouveau. La mémoire contient l'état courant de la machine de Turing, ainsi que le programme. Ce programme se présente sous la forme d'un tableau qui, pour chaque état possible de la machine, définit la séquence d'instruction à réaliser.

On appelle « entrée » d'une machine de Turing l'état initial du ruban avant de lancer le programme, et « sortie » l'état du ruban lorsque la machine s'arrête. Une fonction f est calculable au sens de Turing si elle peut être décrite par une machine de Turing. Tout élément x de son domaine peut alors être représenté par une séquence de caractères et passé en entrée à la machine. Après un nombre fini d'opérations, la machine s'arrêtera sur une sortie représentant $f(x)$.

La machine universelle est une machine de Turing un peu particulière, capable de simuler le comportement de n'importe quelle autre machine de Turing. Elle prend en entrée deux chaînes de caractères m et x , correspondant respectivement à la description d'une machine de Turing M et à une entrée x admissible pour cette machine. La machine universelle écrit alors $M(x)$.

Fonction récursive générale Dans un premier temps, la notion de "fonction calculable" fut approximée par l'idée de *fonction récursive primitive*. Cette dernière était définie récursivement à partir de fonctions à une variable en faisant intervenir différentes manipulations, comme la composition de fonctions ou l'opérateur de succession.

Par la suite, l'on s'est rendu compte que certaines fonctions récursives, comme la fonction d'Ackermann, ne satisfaisaient pas cette définition. L'on introduisit alors la définition de *fonction récursive générale*. Une fonction $f : D \subseteq \mathbb{N} \rightarrow \mathbb{N}$ est *récursive générale* si l'image de tout $n \in D$ peut être calculée par un algorithme. En d'autres termes, s'il existe un nombre fini d'instructions telles que, après un nombre fini d'étapes bien définies, l'image $f(n)$ est obtenue. Lorsque $D = \mathbb{N}$, l'on parle de *fonction récursive partielle*. L'on doit à Kleene la formalisation de cette notion.

Fonction λ -définissable Dans un système formel du premier ordre, il existe des symboles pour désigner les variables, les constantes, les fonctions et les prédicats. Les fonctions ne suivent pas les mêmes règles syntaxiques que les variables : elles ne peuvent par exemple pas être introduites comme arguments dans d'autres fonctions ou prédicats. Cela crée une hiérarchie complexe entre variables. De plus, ce type de langage autorise de parler de

$f(x)$ sans avoir défini la procédure de calcul permettant de trouver l'image.

L'idée de Church est de construire un système ne faisant intervenir qu'un seul type de variable, en plus d'un symbole spécial noté λ . Ce symbole permet de construire des fonctions en définissant explicitement les opérations à effectuer sur les variables.

Formellement, pour toute expression y de ce système, la formule $\lambda x.y$ dénote une fonction qui a tout x fait correspondre y . Dans cette formule, x est une variable liée, tandis que y peut être soit une variable libre, soit une autre formule.

À titre d'illustration, la fonction identité est notée $\lambda x.x$, tandis que les nombres naturels sont construits de la façon suivante (en interprétant y comme représentant le "0" et x comme l'opérateur "successeur") :

$$\begin{aligned} N_1 &\equiv \lambda xy.xy \\ N_2 &\equiv \lambda xy.x(xy) \\ N_3 &\equiv \lambda xy.x(x(xy)) \\ &\dots \end{aligned}$$

Dans ce contexte, une fonction f est dite λ -définissable s'il est possible de la représenter en λ -calcul. Plus formellement, si N_a et N_b représentent des entiers a et b , il doit exister une formule F du lambda calcul tel que :

$$\begin{array}{ccc} f(a) & \xrightarrow{\text{calcul}} & b \\ FN_a & \xrightarrow{\text{conversion}} & N_b \end{array}$$

Où la "conversion" désigne une succession de transformations autorisées par les règles syntaxiques du λ -calcul (la substitution, la réduction et l'expansion).

Théorème de Turing

Kleene a démontré l'équivalence entre les fonctions récursives générales et certains types de fonctions λ -définissables. Turing pour sa part a montré l'équivalence de ces deux notions avec sa propre définition de fonction calculable.

Théorème 9. On a l'équivalence entre²⁵ :

1. Fonction récursive générale.
2. Fonction Turing-calculable.

25. LADRIÈRE, *Ibidem*, p. 236.

3. Les fonctions λ -définissables définies sur tout $\mathbb{N}^n$, où n est le nombre de variables.

Il est possible d'inclure les fonctions λ -définissables dont le domaine de définition est $D^n \subset \mathbb{N}^n$. Dans ce cas, l'équivalence devient :

Théorème 10. On a l'équivalence entre^[26] :

1. Fonction récursive partielle.
2. Fonction λ -définissable.

Chacune de ces trois définitions de la calculabilité peut donc être utilisée pour tenter de résoudre le problème de décision. Church s'est penché sur ce problème avec la notion de λ -définissabilité, Kleene avec celle de récursivité générale et Turing avec la sienne. Ils ont tous les trois obtenu un résultat très analogue.

Théorème 11. Aucune théorie satisfaisant les hypothèses des théorèmes d'incomplétude de Gödel-Rosser n'est résoluble^[27].

Il n'existe donc pas d'algorithme universel permettant de déterminer la démontrabilité d'une proposition dans de telles théories. La méthode de la preuve utilisée dans ce théorème, pour chacune des trois définitions de la calculabilité, utilise un argument diagonal.

26. LADRIÈRE, *Ibidem*.

27. LADRIÈRE, *Ibidem*, p. 238-244.

Annexe 4 : Le paradoxe de Cantor

Le paradoxe de Cantor, prouvé en 1899, stipule qu'il n'existe pas d'ensemble de tous les ensembles; ou, plus rigoureusement, de cardinaux de tous les cardinaux. La notion de cardinal est introduite par Cantor pour hiérarchiser les ensembles en fonctions de leurs tailles, particulièrement les ensembles infinis. Contrairement à ce que l'on pourrait penser, l'infini n'est pas donné une fois pour toute. Il existe des infinis "plus grands" et des infinis "plus petits". L'ensemble $\mathbb{N}$ des entiers naturels, par exemple, est plus petit que l'ensemble $\mathbb{R}$ des nombres réels. Ils comportent pourtant tous les deux un ensemble infini de nombre.

Pour comprendre en quoi exactement on peut dire d'un ensemble infini qu'il est plus grand qu'un autre, je vais commencer par donner trois définitions de "taille" d'un ensemble, y compris la définition de cardinalité utilisée par Cantor. Je vais ensuite montrer comment ces concepts peuvent être utilisés pour montrer que l'ensemble de tous les ensembles n'est jamais de taille assez grande pour contenir l'ensemble de ses sous-ensembles. Cela nous demandera d'introduire un argument connu sous le nom « d'argument de la diagonale. »

Comment définir la taille d'un ensemble infini ?

Il semble naturel de penser que la « taille » d'un ensemble soit donnée par le nombre d'éléments compris dans cet ensemble. L'ensemble $E_1 = \{2, 4, 6, 8\}$ est plus grand que l'ensemble $E_2 = \{2, 4, 8\}$ car il comprend un élément de plus. Selon cette définition, il coule de source que l'ensemble E_1 aura une taille plus grande ou égale que chacun de ses sous-ensembles. Un sous-ensemble de E_1 est en effet obtenu en choisissant un ou plusieurs des éléments de E_1 . On peut avoir, à titre d'exemple, $\{2, 8\}$ $\{4, 6, 8\}$, mais aussi l'ensemble vide $\{\}$ et l'ensemble E_1 lui-même.

En présence d'ensembles infinis, il est difficile de compter le nombre d'éléments dans un ensemble, si bien que la notion de taille devient plus difficile à définir. Pour s'en rendre compte, intéressons-nous aux deux ensembles sui-

vants :

$$\begin{aligned} A &= \{2, 4, 6, 8, 10, \dots\} \\ B &= \{2, 4, 8, 16, 32, \dots\} \end{aligned}$$

La question que l'on se pose est celle de savoir lequel de ces deux ensembles est le plus "grand", et en quel sens exactement ? À prime abord, ils ont tous les deux une infinité d'éléments, mais sont-ils pour autant de même taille ? J'aimerais donner trois approches distinctes et les discuter : une approche utilisant la notion d'inclusion, une approche fondée sur la notion de densité et une autre reposant sur la notion de cardinalité.

Approche 1 : L'Inclusion

On remarque tout d'abord que l'ensemble B des puissances de deux est inclus dans l'ensemble A des nombres pairs, ce qui se note $B \subseteq A$. On peut donc penser que B est "plus petit" que l'ensemble A . Pour formaliser cette notion de "taille", restée pour le moment à l'état d'intuition, on pourrait avancer la définition suivante :

Définition 30. Soient deux ensembles A et B . On pose :

- $A \prec B$ si et seulement si $A \subset B$.
- $A \preceq B$ si et seulement si $A \subseteq B$.
- $A \approx B$ si et seulement si $B \subseteq A \subseteq B$.

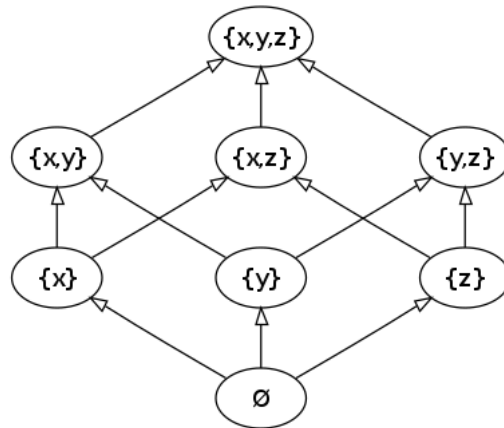
Les symboles $\prec$, $\preceq$ et $\approx$ intervenant dans cette définition doivent se lire respectivement "est plus petit que", "est plus petit ou égal à" et "est égal à". Cette identification du symbole " $\preceq$ " avec le symbole " $\subseteq$ " n'est pas qu'une simple métaphore. On peut montrer que les propriétés algébriques essentielles attendues de la relation "être algébriquement plus petit ou égal à" (" $\preceq$ ") sont satisfaites par la relation "être un sous ensemble de" (" $\subseteq$ "). On a en effet :

$$\begin{aligned} A &\preceq A && \text{(réflexivité)} \\ \text{si } A &\preceq B \text{ et } B \preceq A, \text{ alors } A \approx B && \text{(antisymétrie)} \\ \text{si } A &\preceq B \text{ et } B \preceq C, \text{ alors } A \preceq C && \text{(transitivité)} \end{aligned}$$

Mais on a aussi :

$$\begin{aligned} A &\subseteq A && \text{(réflexivité)} \\ \text{si } A &\subseteq B \text{ et } B \subseteq A, \text{ alors } A = B && \text{(antisymétrie)} \\ \text{si } A &\subseteq B \text{ et } B \subseteq C, \text{ alors } A \subseteq C && \text{(transitivité)} \end{aligned}$$

Ces trois propriétés définissent ce que l'on appelle, en logique mathématique, un *ordre partiel* sur la classe des ensembles. Elles nous permettent de comparer des ensembles entre eux et de les hiérarchiser. J'en donne une illustration dans le cas d'ensembles comportant trois éléments x, y et z .



Le diagramme ??, appelé diagramme de Hasse, donne une hiérarchisation des sous-ensembles de $\{x, y, z\}$ en utilisant la relation d'ordre " $\subseteq$ ". L'ensemble vide " $\emptyset$ " se retrouve tout en-dessous car il est un sous-ensemble de tous les autres ensembles, tandis que $\{x, y, z\}$, qui n'est contenu dans aucun des autres ensembles de $\mathbb{P}(\{x, y, z\})$, se retrouve tout au-dessus. On observe que les ensembles $\{x\}$, $\{y\}$ et $\{z\}$, qui n'ont chacun qu'un seul élément, se trouvent au premier niveau. De la même manière, les ensembles $\{x, y\}$, $\{x, z\}$ et $\{y, z\}$, à deux éléments chacun, se retrouvent au deuxième niveau. Il n'y a pas de flèches entre les ensembles d'un même niveau car ces ensembles sont incomparables les uns les autres. En effet, on a $\{x\} \not\subseteq \{y\}$ et $\{y\} \not\subseteq \{x\}$.

On voit dans l'exemple ci-dessus que tous les ensembles ne sont pas comparables les uns les autres. Cela peut arriver même en présence d'ensembles qui ne contiennent pas le même nombre d'éléments. À titre d'exemple, l'ensemble $E = \{1, 2, 3\}$ et $E' = \{1, 3, 4, 5\}$ ne sont pas comparables car $E \not\subseteq E'$ et $E' \not\subseteq E$. Pourtant, on aimerait pouvoir dire que l'ensemble E , dans la mesure où il contient plus d'éléments, est plus grand que E' . Il faut donc évoluer vers une autre définition.

Approche 2 : La Densité

Pluôt que de se demander si un ensemble est inclus dans un autre, ce qui représente finalement un cas assez exceptionnel, on pourrait raisonner de la façon suivante : les deux ensembles A et B possèdent certes une infinité de nombres entiers, mais il y a également une infinité d'entiers qu'ils ne possèdent pas. L'ensemble A des nombres pairs possède exactement la moitié des

entiers naturels, vu qu'il lui manque les nombres impairs. L'ensemble B des puissances de deux ne contient qu'une fraction de nombres pairs ; non seulement il lui manque les nombres impairs, mais aussi la plupart des nombres pairs qui ne sont pas des puissances de deux, comme 6, 10, 12, 14, 18, 20, ... On aimerait donc dire que l'ensemble A est plus *dense* que l'ensemble B .

Pour formaliser cette notion de *densité*, demandons-nous quelle est la probabilité, si l'on prenait un entier au hasard, que cet entier soit dans A ou qu'il soit dans B ? Supposons, pour commencer avec un cas évident, que le nombre entier pris au hasard soit compris entre 0 et 100. Il est sélectionné d'après une loi de probabilité homogène, ce qui signifie qu'il y a exactement une chance sur cent de tomber sur 0, une chance sur cent de tomber sur 1, etc.

$$A = \{ \underbrace{2, 4, 6, \dots, 98}_{49 \text{ nombres pairs}}, 100, 102, \dots \}$$

$$B = \{ \underbrace{2, 4, 8, 16, 32, 64}_{6 \text{ puissances de deux}}, 128, 256, \dots \}$$

Sur l'intervalle $[0, 99]$, on compte exactement 49 nombres pairs non nuls et 6 puissances de deux différentes de un. Il y a donc plus ou moins une chance sur deux que le nombre soit dans A et 5% de chance qu'il soit dans B . Cela confirme l'idée que A est plus dense que B . Il est toutefois arbitraire de prendre un entier sur l'intervalle $[0, 99]$ plutôt que sur l'intervalle $[0, 999]$ ou $[0, 9999]$. C'est pourquoi on aimerait savoir que deviennent ces probabilités si l'intervalle est $[0, n]$ lorsque la borne n de cet intervalle tend vers l'infini ?

Si on augmente la taille de l'intervalle $[0, 99]$, on constate que le nombre de naturels pairs augmente de façon linéaire, tandis que le nombre de puissance de deux suit une courbe décroissante de type logarithmique. Notez que, dans les tables ci-dessous, la notation $\lfloor \cdot \rfloor$ désigne l'opération d'arrondir à l'unité inférieure (ex : $\lfloor \log_2(3) \rfloor = \lfloor 1,58496 \rfloor = 1$).

Intervalle	$[1, 100]$	$[1, 200]$	$[1, 300]$	...	$[1, n]$
# d'entiers dans B	50	100	150	...	$\lfloor n/2 \rfloor$
Probabilité	50%	50%	50%	...	50%

Intervalle	$[1, 100]$	$[1, 200]$	$[1, 300]$	...	$[1, n]$
# d'entiers dans A	6	7	8	...	$\lfloor \log_2(n) \rfloor$
Probabilité	6%	3.5%	2.6%	...	?

Pour arriver à un résultat précis, on cherche les formules P_A et P_B donnant la probabilité qu'un entier tiré au hasard sur l'intervalle $[0, n]$ soit respectivement dans l'ensemble A et B . Cette probabilité est donnée par :

$$P_A(n) = \frac{\# \text{ de naturels pairs dans l'intervalle}}{\text{taille de l'intervalle}} \\ = \frac{\lfloor \frac{n}{2} \rfloor}{n}$$

et

$$P_B(n) = \frac{\# \text{ de puissance de deux dans l'intervalle}}{\text{taille de l'intervalle}} \\ = \frac{\lfloor \log_2(n) \rfloor}{n}$$

Il est facile de constater que lorsque la taille de l'intervalle $[0, n]$ tend vers l'infini, la probabilité $P_B(n)$ tend vers $1/2$ tandis que la probabilité $P_A(n)$ tend vers zéro. Il n'y a donc pratiquement aucune chance qu'un naturel soit une puissance de deux en général. Cette observation nous permet de conclure que l'ensemble A est infiniment plus dense que l'ensemble B .

On pourra formaliser la notion de "taille" entre deux ensembles A et B quelconque de la manière suivante :

Définition 31. Soient deux ensembles A et B . On pose :

- $A \prec_\delta B$ si et seulement $\lim_{n \rightarrow \infty} P_A(n) < \lim_{n \rightarrow \infty} P_B(n)$.
- $A \preceq_\delta B$ si et seulement $\lim_{n \rightarrow \infty} P_A(n) \leq \lim_{n \rightarrow \infty} P_B(n)$.
- $A \approx_\delta B$ si et seulement si $\lim_{n \rightarrow \infty} P_A(n) = \lim_{n \rightarrow \infty} P_B(n)$.

J'ai utilisé la notation $\prec_\delta$, avec un petit δ en indice, pour différencier cette approche de celle utilisée précédemment, qui reposait sur la notion d'inclusion. Il est possible de montrer que $\prec_\delta$ forme une *relation d'ordre* et respecte donc les propriétés algébriques de *réflexivité*, *d'antisymétrie* et de *transitivité* qui ont été vues précédemment.

L'avantage de cette nouvelle définition de "taille", c'est qu'elle permet de comparer entre eux des ensembles qui n'étaient pas comparables en utilisant l'inclusion (" $\subseteq$ "). Si on reprend par exemple l'ensemble des parties de $\{x, y, z\}$, est-il possible de comparer $\{x, y\}$ et $\{x, z\}$ en termes de densité ? Oui car la probabilité qu'un élément e prit au hasard dans $\{x, y, z\}$ soit dans $\{x, y\}$ est égale à la probabilité que cet élément soit dans $\{x, z\}$. En termes de densité, les deux ensembles ont donc la même taille, bien qu'ils ne soient pas inclus l'un dans l'autre. Cependant, on va voir que ce n'est pas la seule mesure admissible de la "taille" d'un ensemble.

Un défaut de cette approche par densité vient de ce que, pour être comparables, les ensembles A et B doivent être tous deux des sous-ensembles

d'un même ensemble commun. Dans notre exemple, cet ensemble commun est $\mathbb{N}$. On aurait pu aussi choisir $\mathbb{R}$ ou $\mathbb{C}$, et on aurait alors obtenu des densités différentes (la probabilité qu'un réel $r \in \mathbb{R}$ tiré au hasard soit un naturel pair est égale à zéro).

On peut toutefois se demander ce qu'il se passe si les ensembles A et B n'appartiennent pas à un ensemble commun ? Dans ce cas, ils ne sont tous simplement pas comparables avec la méthode des densités. Pour illustrer cette situation, considérons l'ensemble $A' = \{a_1, a_2, \dots, a_n\}$ et $B' = \{b_1, b_2, \dots, b_m\}$, où les $a_1, \dots, a_n$ sont des lettres d'un alphabet $\mathcal{X}_a$ (par exemple l'alphabet romain) et les $b_1, \dots, b_m$ sont des lettres d'un alphabet $\mathcal{X}_b$ distinct de $\mathcal{X}_a$ (par exemple l'alphabet russe). Il est toutefois possible de contourner cette difficulté en constatant que A' et B' sont tous les deux dans $\mathcal{X}_a \cup \mathcal{X}_b$.

Approche 3 : La Cardinalité

Je vais maintenant montrer qu'on peut dériver un argument contredisant le résultat obtenus à l'aide des deux méthodes précédentes, qui voulaient que l'ensemble A des nombres pairs et l'ensemble B des puissances de deux auraient la même taille. Mon argument consiste à s'intéresser, non pas à l'inclusion ou à la densité, mais à ce qu'on appelle la cardinalité.

La cardinalité consiste à raisonner de la façon suivante. Supposons que l'on possède deux grands sac remplis d'objets. Pour savoir lequel des deux est le plus grand, ôtant un objet au hasard du premier sac, puis ôtant-en un autre au hasard du second sac. Le sac qui est vide en premier est le plus petit. S'ils se vident simultanément, ils ont la même taille. Que devient cette approche dans le cas de nos ensembles infinis A et B ? Intuitivement, on aurait tendance à penser qu'ils ont la même taille, puisqu'ils ne se videront jamais ni l'un ni l'autre ? Comment montrer cette propriété de façon plus rigoureuse ? Comment déterminer les cas particuliers où un ensemble infini se "vide" plus vite qu'un autre ?

Remarquons qu'il est possible de mettre en correspondance l'ensemble A avec l'ensemble des entiers naturels $\mathbb{N}$

$$\begin{array}{ccccccc} \mathbb{N} : & 0 & 1 & 2 & 3 & \dots & n \\ & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow \\ A : & 2 & 4 & 6 & 8 & \dots & 2(n+1) \end{array}$$

Ce tableau suggère qu'il est possible d'associer, à tout entier $n \in \mathbb{N}$, un unique élément $a \in A$, et qu'inversement il est possible de faire correspondre, à chaque élément $a \in A$, un unique entier $n \in \mathbb{N}$. Ce genre de relation est ce que l'on appelle une *bijection* de A dans $\mathbb{N}$.

Intuitivement, la présence d'une bijection signifie qu'il y a "autant" d'entiers dans $\mathbb{N}$ que d'éléments dans A : chaque fois que l'on retire un élément de l'ensemble des entiers, on peut en retirer un de l'ensemble A ; et inversement. Si bien que, de ce point de vue, les deux ensembles ont la même

"taille". En termes techniques, on dit que A et $\mathbb{N}$ sont **équinumérables**, ou encore, qu'ils ont la même **cardinalité**.

La question que l'on se pose maintenant est celle de savoir si B et $\mathbb{N}$ sont eux aussi équinumérables ? La réponse est positive. Pour s'en rendre compte, il suffit de dresser la correspondance suivante :

$$\begin{array}{ccccccc} \mathbb{N} : & 0 & 1 & 2 & 3 & \dots & n \\ & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow \\ B : & 2 & 4 & 8 & 16 & \dots & 2^{n+1} \end{array}$$

Dans la mesure où A est équinumérable à $\mathbb{N}$ et où $\mathbb{N}$ est équinumérable à B , on peut montrer que A est aussi équinumérable à B . Il suffit de prendre la bijection ci-dessous :

$$\begin{array}{ccccccc} \mathbb{N} : & 2 & 4 & 6 & 8 & \dots & 2n \\ & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow \\ B : & 2 & 4 & 8 & 16 & \dots & 2^n \end{array}$$

Ces observations nous suggèrent que A et B ont la même "taille" (en terme de cardinalité), bien que (1) il y a infiniment plus de nombres pairs que de puissances de deux (en termes de densité) et que (2) l'ensemble B soit un sous-ensemble de A . Dans la notation de Cantor, on dira que A , B et $\mathbb{N}$ sont équipotents et on désignera leur taille à l'aide du symbole $\aleph_0$.

Au vu de ces observations, on est en droit de se demander s'il existe des ensembles qui ne sont pas équipotents à $\mathbb{N}$? Le théorème de Cantor, qu'on verra plus loin, répond par l'affirmative et montre que :

- L'ensemble $\mathbb{R}$ a une cardinalité strictement supérieure à la cardinalité de $\mathbb{N}$, ce que l'on note $|\mathbb{R}| > \aleph_0$;
- L'ensemble des parties de $\mathbb{R}$ a une cardinalité strictement supérieure à la cardinalité de $\mathbb{R}$, ce que l'on note $|\mathcal{P}(\mathbb{R})| > |\mathbb{R}|$;
- L'ensemble des parties des parties de $\mathbb{R}$ a une cardinalité strictement supérieure à la cardinalité des parties de $\mathbb{R}$, ce que l'on note $|\mathcal{P}(\mathcal{P}(\mathbb{R}))| > |\mathcal{P}(\mathbb{R})|$;
- etc.

À ce stade, un lecteur scrupuleux demandera peut-être ce que cela signifie pour un ensemble d'avoir une cardinalité strictement plus grande que celle d'un autre ensemble ? Jusqu'à présent en effet, nous avons défini ce que ça voulait dire pour deux ensembles d'avoir la même cardinalité, mais nous n'avons pas encore défini de relation d'ordre.

Pour répondre à cette question, considérons les ensembles $A' = \{x, y\}$ et $B' = \{x, y, z\}$. On remarque qu'il est impossible d'établir une bijection entre A' et B' et qu'ils n'ont donc pas la même cardinalité. En revanche, il est possible de faire une *injection totale* de A' dans B' et une *surjection* de

B' dans A' . C'est en raison de ces relations que l'on dira que la cardinalité de B' est plus grande que celle de A' .

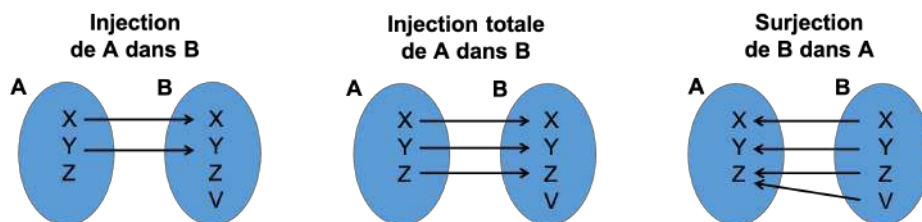


FIGURE 16.3 – La figure de gauche illustre une injection de A dans B . Cette transformation est telle qu'il n'existe pas deux éléments de A qui ont la même image. La figure du milieu est une injection totale, ce qui veut dire que non seulement la transformation est injective, mais qu'en plus de ça tous les éléments de A ont une image dans B (on écrira $A \preceq_\alpha B$). Le graphique de droite, pour finir, illustre une surjection de B' dans A' . Cette transformation est telle que tout élément de A' est une image d'un élément de B' (on écrira $B \succeq_\alpha A$). Notons que, pour la figure de gauche, on ne peut pas comparer A et B car certains éléments de A n'ont pas d'image dans B et que certains éléments de B ne sont pas des images d'éléments de A (on ne pourra ni écrire $A \preceq_\alpha B$ ni $B \succeq_\alpha A$ ni $A \approx_\alpha B$).

Cela suggère la définition suivante :

Définition 32. Soient deux ensembles A et B . On pose :

- $A \approx_\alpha B$ ssi il existe une bijection de A dans B .
- $A \preceq_\alpha B$ ssi il existe une injection totale de A dans B .
- $A \succeq_\alpha B$ ssi il existe une surjection de A dans B .

Les symboles $\approx_\alpha$, $\preceq_\alpha$ et $\succeq_\alpha$, où un petit " α " a été ajouté en indice, forment encore une fois une *relation d'ordre* et respectent donc les propriétés de *réflexivité*, d'*antisymétrie* et de *transitivité*^[28].

Dire de la relation $\preceq_\alpha$ qu'elle forme une relation d'ordre implique en particulier que si $A \preceq_\alpha B$ et $B \preceq_\alpha A$, alors $A \approx_\alpha B$. En français dans le texte : s'il existe une injection totale de A dans B et qu'il existe une injection totale de B dans A , alors A et B ont la même cardinalité. La preuve de cet important théorème est donnée par Bernstein et Cantor aux alentours de 1896^[29].

28. L'antisymétrie exprime que si une injection totale existe entre A et B (ce que l'on note $A \preceq_\alpha B$) et si une injection totale existe aussi entre B et A (notée $B \preceq_\alpha A$), alors il existe une bijection entre A et B (notée $A \approx_\alpha B$). Ce théorème, prouvé par Cantor et Bernstein, est loin d'être trivial, mais il est indispensable pour faire de l'équénommabilité une mesure valide de la taille d'un ensemble.

29. Georg Cantor énonce ce théorème sans démonstration en 1887, puis en donne une preuve dans son livre *Sur les fondements de la théorie des ensembles transfinis* (1895-

Propriétés contre-intuitives de l'infini

À l'issue de ces trois approches, celle de l'inclusion, de la densité et de la bijection, l'on se rend compte qu'un certain nombre de propriétés qui paraissent intuitives dans un monde fini doivent être rejetées si l'on souhaite rester cohérent dans en présence d'ensembles infinis :

- **La taille se dit en plusieurs sens** : Il y a plusieurs manières de définir la "taille" d'un ensemble. Une première définition s'appuie sur la notion d'inclusion, une seconde s'exprime en termes de densité et une troisième utilise le concept d'équipotence. En présence d'ensembles finis, ces trois définitions coïncident et peuvent être substituées l'une à l'autre. Ce n'est plus le cas lorsque l'on accepte *l'infini en acte*.
- **Le tout n'est pas plus grand que les parties** : Un principe méréologique d'apparence très intuitive, et fermement défendu par Aristote, énonce que le tout est plus grand que les parties. Bolzano le premier suggère de révoquer ce principe afin de lever certaines difficultés concernant l'existence d'ensembles *infinis en acte*. Cette idée sera confirmée par le développement ultérieur des mathématiques. Dans le cadre spécifique de la théorie de Cantor, il est en effet possible de dresser une bijection entre un ensemble infini et une partie infinie de cet ensemble, comme on l'a vu avec l'exemple des ensembles A et B . Cela montre que la *cardinalité* est différente de *l'inclusion*.
- **Un évènement de probabilité nulle n'est pas impossible** : Pour s'en rendre compte, il suffit de se poser la question suivante : quelle est la probabilité, si l'on tire un entier naturel au hasard (de façon équiprobable), que celui-ci soit une puissance de deux ? Le calcul effectué à l'exercice (c) nous montre que la probabilité qu'un entier m pris aléatoirement dans $\{1, 2, \dots, n\}$ soit une puissance de deux tend vers zéro lorsque n tend vers l'infini. On conclut donc que la probabilité qu'un entier m pris dans $\mathbb{N}$ soit une puissance de deux est nulle. Cela peut paraître difficile à se représenter de prime abord. Comment penser en effet que m ne puisse pas être égal à 2, à 4, à 8, à 16, etc. ? Après tout, ces nombres ne sont-ils pas des candidats admissibles ? Pour se défaire de cet apparent paradoxe, peut-être dira-t-on que la probabilité n'est pas *nulle*, mais *infinitésimale* ? Dire cela, c'est considérer que $\mathbb{N}$ forme un *infini en puissance*. Cependant, l'idée de Cantor est d'assigner à $\mathbb{N}$, conçu comme un tout, une taille déterminée, qu'il note $\aleph_0$. Ce que ce geste a d'inaugural, c'est qu'au lieu de

1897), en utilisant la théorie des ordinaux. Felix Bernstein, élève de celui-ci, en esquisse une autre démonstration dès 1896 à l'âge de 18 ans. Elle est publiée en 1898 sur proposition de Cantor dans *Leçons sur la théorie des fonctions* sous la plume d'Émile Borel.

voir $\mathbb{N}$ comme l'ensemble obtenu de façon constructive (et récursive) en prenant $\{1\} \cup \{2\} \cup \dots \cup \{n\} \cup \dots$, on le voit comme un tout donné en une seule fois. Au lieu de le voir comme le produit d'une énumération infinie du type $\lim_{n \rightarrow \infty} \{1, 2, \dots, n\}$, on le voit comme un authentique *infini en acte*.

Or dès lors que $\mathbb{N}$ forme un infini en acte, la probabilité qu'un entier naturel m tiré au hasard (de façon équiprobable) soit une puissance de deux n'est pas infinitésimale, mais elle est rigoureusement égale à zéro. Il faut donc distinguer la possibilité mathématique de la possibilité probabiliste. Expliquons-nous. Un événement mathématiquement contradictoire (comme tirer un entier naturel à la fois pair et impair) a toujours une probabilité nulle. En revanche, un événement non contradictoire (comme tirer un entier naturel qui soit une puissance de deux) n'a pas toujours une probabilité non nulle. Pour conclure, on peut dire qu'accepter l'existence d'un infini en acte, c'est reconnaître *de facto* une séparation entre le règne des possibilités et le règne des probabilités.

Si la distinction entre "impossibilité logique" et "impossibilité probabiliste" n'est toujours pas claire à vos yeux, je vous invite à participer à un petit jeu. Je vais penser à un nombre réel entre 0 et 1. Ensuite, je vais vous demander de me donner un nombre de votre choix. Si ce nombre est le même que celui que j'avais à l'esprit, vous recevrez 20/20 à l'examen. Sinon, rien ne se passera. La question est la suivante : avez-vous intérêt à accepter le jeu ? On est tenté de dire oui, car dans le fond vous n'avez rien à perdre. Pourtant, vos chances de gains sont strictement nulles. Pourquoi cela ? Parce qu'un nombre réel est construit avec une infinité de décimales. Pour peu que vous ayez deviné juste les dix premières décimales du nombre que j'avais en tête, quelles sont les chances que vous deviniez les dix suivantes ? Les cent suivantes ? Les mille suivantes ? Les $\aleph_0$ suivantes ?

L'argument de la diagonale de Cantor

Maintenant que nous avons clarifié quelques particularités des ensembles infinis, j'aimerais introduire l'argument de la diagonale de Cantor. Cet argument est un ingrédient clef dans la preuve d'après laquelle il n'existe pas d'ensemble de tous les ensembles.

L'Ensemble des sous-ensembles

L'argument de la diagonale consiste à montrer que tout ensemble E a une taille strictement plus petite que l'ensemble de ses sous-ensembles $\mathbb{P}(E)$. Pour rappel, l'ensemble des sous-ensembles est obtenu en énumérant tous les

sous-ensembles qu'il est possible de construire avec un ensemble de départ. L'ensemble $\{1, 2\}$, par exemple, admet quatre sous-ensembles : $\{\}$, $\{1\}$, $\{2\}$, $\{1, 2\}$. Il est facile de montrer que si l'ensemble E compte n entiers, l'ensemble $\mathbb{P}(E)$ en comptera 2^n .

La question est de savoir que devient la taille de $\mathbb{P}(E)$ si E est un ensemble infini ? L'argument de la diagonale répond que l'ensemble $\mathbb{P}(E)$ est toujours plus grand, en cardinalité, que l'ensemble de départ E . Cela veut dire que si on utilise la notion de taille formalisée par l'usage de la bijection, $E \prec_{\alpha} \mathbb{P}(E)$.

Pour répondre à cette question, supposons que E et $\mathbb{E}$ ait la même taille. Cela signifie qu'on pourrait créer un livre, doté d'une infinité dénombrable de pages, dont chaque page comporterait un ensemble de naturels distincts. La page une comporterait par exemple l'ensemble $\{1, \{2, 3\}\}$, la page deux comporterait l'ensemble $\{\{7\}\}$, etc. L'on suppose que toutes les pages sont distinctes les unes des autres, mais qu'il y en a une infinité dénombrable, si bien que le livre, considéré dans son intégralité, dénombre tous les ensembles possibles de naturels.

Ce livre dont je parle est une bijection entre l'ensemble des entiers et l'ensemble des sous-ensembles. Chaque page du livre correspond en effet à un entier : on a la page portant le numéro 1, la page portant le numéro 2, etc. De plus, il est écrit sur chaque page un sous-ensemble d'entiers. Il y a donc bien une correspondance biunivoque entre N et $\mathbb{P}(N)$.

Cela étant posé, j'aimerais montrer que ce livre est contradictoire et qu'il ne peut pas exister. Pour ce faire, appelons "nombre ordinaire" un entier n qui n'est pas représenté sur la page portant le numéro n . Le nombre 2 est un nombre ordinaire, puisque l'ensemble représenté sur la page numéro 2, à savoir $\{\{7\}\}$, ne comporte pas le nombre 2. Le nombre 1, en revanche, n'est pas un nombre ordinaire dans notre exemple, puisque la page numéro 1 contient le nombre 1 dans l'ensemble qu'elle représente : $\{1, \{2, 3\}\}$.

Comme le livre de Cantor comporte tous les ensembles de naturels, il doit exister une page, portant le numéro n' , contenant l'ensemble de tous les nombres ordinaires, et aucun autre. L'ensemble représenté sur la page n' devra donc, entre autres, contenir le nombre 2, mais pas le nombre 1. Vous me suivez ? On pose dès la question suivante : n' est-il un nombre ordinaire ?

Supposons que n' soit un nombre ordinaire. Il doit donc être représenté sur la page qui contient tous les nombres ordinaires. Or cette page porte le numéro n' . Cela contredit la définition de nombre ordinaire. Supposons que n' ne soit pas un nombre ordinaire. Par définition, cela veut dire qu'il doit être représenté sur la page portant le numéro n' . Cependant, cette page ne comporte que les nombres ordinaires, et aucun autre. Cela contredit le fait que n' soit un nombre ordinaire. On en conclut que le livre de Cantor est paradoxal.

Si on veut éviter que les mathématiques se contredisent, il faut refuser l'existence d'un tel livre. Qu'est-ce que cela implique-t-il d'un point de vue

logique ? Qu'il est impossible d'énumérer tous les ensembles de nombres naturels en les consignant dans un livre infini. D'un point de vue formel, on peut dire qu'il n'existe pas de *bijection* de $\mathbb{N}$ dans $\mathcal{P}(\mathbb{N})$. En revanche, on peut créer une *injection totale* de $\mathbb{N}$ dans $\mathcal{P}(\mathbb{N})$. Pour ce faire, il suffit d'associer chaque naturel n à un ensemble de naturels de telle manière que deux naturels différents n_1 et n_2 n'aient pas les mêmes images. Un exemple d'injection totale est donné ci-dessous :

$$\begin{array}{ccccccccc}
 \mathbb{N} : & 1 & 2 & 3 & \dots & n & & \text{(numéros de pages)} \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow & \\
 \mathcal{P}(\mathbb{N}) : & \{1\} & \{2\} & \{3\} & \dots & \{n\} & & \text{(ensembles distincts de naturels)}
 \end{array}$$

Cet exemple correspondrait à un livre infini dont chaque page de numéro n présenterait l'unique singleton $\{n\}$. On remarque que, dans ce livre, il n'y aurait aucun nombre ordinaire. De plus, aucune page ne comporterait l'ensemble des nombres ordinaires. Il n'y aurait donc pas de paradoxe. Cela veut dire que notre injection totale de $\mathbb{N}$ dans $\mathcal{P}(\mathbb{N})$ est bien fondée et que $\mathbb{N} \preceq_{\alpha} \mathcal{P}(\mathbb{N})$. Comme on sait en outre qu'il n'y a pas de bijection de $\mathbb{N}$ dans $\mathcal{P}(\mathbb{N})$, on peut conclure que l'inégalité est stricte : $\mathbb{N} \prec_{\alpha} \mathcal{P}(\mathbb{N})$.

Généralisation à n'importe quel ensemble

Le raisonnement que nous venons de produire ici est généralisable à n'importe quel ensemble E , que E soit un ensemble fini ou infini, qu'il comporte des nombres, des lettres ou des objets quelconques. Il reste même valide pour des ensembles infinis de cardinalité plus grande que $\mathbb{N}$, comme par exemple $\mathbb{R}$ ou $\mathcal{P}(\mathbb{R})$. Je vais donner ici un schéma de la preuve comme quoi $E \prec_{\alpha} \mathcal{P}(E)$ pour tout ensemble E . La démarche se fait en deux temps. D'abord, on commence à montrer qu'il existe une *injection totale* de E dans $\mathcal{P}(E)$:

$$\begin{array}{ccccccccc}
 E : & x_1 & x_2 & x_3 & x_4 & \dots & x_n & \text{(éléments de E)} \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow & \\
 \mathcal{P}(E) : & \{x_1\} & \{x_2\} & \{x_3\} & \{x_4\} & \dots & \{x_n\} & \text{(ensembles de E)}
 \end{array}$$

Cela nous permet de dire que $E \preceq_{\alpha} \mathcal{P}(E)$. Pour montrer que l'inégalité est stricte, il reste encore à montrer qu'il n'existe pas de *bijection* de E dans $\mathcal{P}(E)$. Pour ce faire, on va supposer qu'il existe une telle bijection et on va montrer que cela conduit à une contradiction. Notons cette bijection f :

$$\begin{array}{ccccccccc}
 E : & x_1 & x_2 & x_3 & x_4 & \dots & x_n & \text{(éléments de E)} \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow & \\
 \mathcal{P}(E) : & f(x_1) & f(x_2) & f(x_3) & f(x_4) & \dots & f(x_n) & \text{(images)}
 \end{array}$$

On considère l'ensemble $C = \{x \in E \mid x \notin f(x)\}$. En français dans le texte : l'ensemble C est l'ensemble des éléments $x \in E$ tels que x n'est pas contenu dans l'ensemble avec lequel il est mis en bijection. Dans le cas particulier où $E = \mathbb{N}$, l'ensemble C ne désigne en fait rien d'autre que l'ensemble des nombres ordinaires.

On remarque que C est un sous-ensemble de E . Il existe donc un $c \in E$ tel que $f(c) = C$ (car on a supposé que f est une bijection de E dans $\mathcal{P}(E)$). On pose alors la question suivante : c est-il contenu dans C ? Si $c \in C$, cela signifie que $c \notin f(c)$ (par définition de C), ce qui implique la contradiction $c \notin C$ (car $f(c) = C$). Mais si $c \notin C$, cela veut dire que $c \in f(c)$ (par définition de C) et donc on a la contradiction $c \in C$ (car $f(c) = C$). On conclut que C est paradoxal et que la bijection f ne peut pas exister. Par conséquent, on a $E \prec_\alpha \mathcal{P}(E)$.

Pour résumer tout ce que je viens de dire en des termes simples, supposons que vous ayez une boîte E contenant toutes les pralines du monde : vous avez des pralines bleu, des rouges, des jaunes, etc. Vous vous demandez combien de boîtes de pralines différentes il est possible de concevoir en piochant des pralines dans cette boîte initiale ? Vous pouvez, par exemple, créer une boîte de pralines ne contenant que le bleu marine et le rouge opalescent, et exclure toutes les autres couleurs ; vous pouvez créer une boîte avec toutes les pralines foncées, toutes les pralines clairs, toutes les pralines bleues, etc. Vous pouvez aussi créer des boîtes de pralines contenant d'autres boîtes de pralines, qui à leur tour contiennent d'autres boîtes de pralines, et ainsi de suite. L'argument de la diagonale de Cantor peut être utilisé pour montrer que la boîte contenant toutes les boîtes de pralines possibles et imaginables sera si grand qu'elle ne rentrera jamais dans la boîte E contenant toutes les pralines du monde en un exemplaire. C'est ce que nous enseigne le théorème de Cantor.

Le paradoxe de Cantor

J'ai montré que pour tout ensemble E donné, l'ensemble $\mathcal{P}(E)$ a une taille strictement plus grande que E . Appliquons ce théorème pour montrer que l'ensemble de tous les ensembles, noté ENS , est paradoxale. On remarque tout d'abord que comme ENS est l'ensemble de tous les ensembles, il contient l'ensemble $\mathcal{P}(ENS)$. On a donc $\mathcal{P}(ENS) \in ENS$. D'autre part, ENS contient aussi l'ensemble des sous-ensembles de ENS , noté $\mathcal{P}(ENS)$. On peut donc trivialement construire une injection totale de $\mathcal{P}(ENS)$ dans ENS :

$$\begin{array}{ccccccc}
 \mathcal{P}(ENS) : & x_1 & x_2 & x_3 & x_4 & \dots & x_n & (\text{éléments de } \mathcal{P}(ENS)) \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \dots & \downarrow & \\
 ENS : & x_1 & x_2 & x_3 & x_4 & \dots & x_n & (\text{images})
 \end{array}$$

Cela nous permet de conclure que $\mathcal{P}(ENS) \prec_{\alpha} ENS$. Or au vu du théorème de Cantor, on peut aussi écrire que $ENS \prec_{\alpha} \mathcal{P}(ENS)$, ce qui conduit à un paradoxe. Si on veut préserver la non-contradiction des mathématiques, on est obligé de refuser l'existence de l'ensemble ENS . Ce résultat est assez déconcertant, car on ne voit pas *à priori* la raison pour laquelle il nous serait interdit de parler de l'ensemble contenant tous les ensembles possibles.

À l'époque de la publication du paradoxe de Cantor (1899), les mathématiciens ne prennent pas au sérieux la menace que représente ce théorème pour le projet d'axiomatisation des mathématiques. En effet, la théorie de Cantor est vue comme douteuse, voire dangereuse, pour cette raison qu'elle admet de parler d'ensemble infinis en acte. Plutôt que de chercher à reconstruire la théorie des ensembles, il est plus facile de refuser en bloc la théorie des ordinaux et des cardinaux. Des personnes de renom, comme Hilbert, diront qu'il faut donner aux mathématiques une approche rigoureuse, où chaque objet mathématique est construit de façon explicite à partir d'un nombre fini d'opérations. Cette exigence exclut *de facto* la possibilité de parler de $\mathbb{N}$ comme d'un ensemble infini en acte. Tout ce dont il est possible de parler, c'est de l'ensemble $\{1, 2, 3, \dots, n\}$, où n est un naturel fini aussi grand que souhaité.

Plutôt que d'insister sur l'existence des objets abstraits, et potentiellement contradictoires, Hilbert considère le *signe* comme la seule définition admissible d'un objet mathématique. Loin d'être une idéalité platonicienne, loin d'être une intuition pure, ce signe est tout ce qui est des plus palpables : il existe sur une feuille de papier, suit des règles de syntaxes très strictes et toutes les manipulations dont il peut faire l'objet sont visibles, traçables, vérifiables. Bien sûr, Hilbert était conscient que, pour faire des mathématiques, il est parfois nécessaire d'utiliser des abréviations ou des raisonnements pas tout à fait formalisés.

C'est pourquoi il distingue la métamathématique, dont le rôle est de montrer la consistance des opérations mathématiques usuelles, des mathématiques proprement dites. Au sein de la métamathématique, tout doit pouvoir être dérivé des axiomes à l'aide d'un nombre fini de signes. Hilbert imagine à ce titre la fiction d'un mathématicien disposant d'un crayon et d'un bloc de feuilles de papier de taille finie, mais rendue aussi grande que souhaitée. Ce mathématicien pourra-t-il dériver tous les résultats mathématiques connus à l'aide d'un nombre fini de signes ? Pourra-t-il déterminer la solution aux 23 problèmes mathématiques encore irrésolus à l'aube du XX^{ème} siècle ?

C'est le défi lancé par Hilbert à la communauté scientifique lors du congrès international des mathématiciens de 1900, qui réunit à Paris les experts de toutes les disciplines. Il n'est pas anodin de remarquer combien la figure de ce calculateur envisagé par Hilbert préfigure celle de l'ordinateur. D'ailleurs, le terme *computer*, en anglais, désignera pendant longtemps un homme capable de dériver des calculs sur le papier. Ce n'est qu'après la mise

aux points des premiers ordinateurs que le terme changera progressivement de signification, pour devenir ce que l'on connaît aujourd'hui.

Dans ce climat effervescent, où le formalisme de Hilbert et la volonté d'axiomatiser les mathématiques dominant largement le paysage épistémique, il faudra attendre le paradoxe de Russell, en 1902, pour que soit reconnue par tous la fragilité de la théorie naïve des ensembles. La force de ce nouveau paradoxe, c'est qu'il ne présuppose rien quant à l'existence des ordinaux ou des cardinaux. Il ébranle aussi bien les *Fondements de l'Arithmétique* de Frege que le rêve de Hilbert, et sera à l'origine de nouvelles approches axiomatiques : la théorie des types de Russell, la théorie de Zermelo-Fraenkel et l'ontologie sans "classe" de Lesniewski.

Annexes 5 : Les Nombres transfinis

La définition des nombres ordinaux utilisée est celle de von Neumann, plus facile à formaliser dans la théorie des ensembles. Un ordinal α est un ensemble transitif pour lequel la relation d'appartenance $\in$ définit une relation d'ordre stricte. Si $x \in \alpha$, alors $x \subset \alpha$ et si $\{x, y\} \in \alpha$, alors soit $x \in y$, soit $x = y$, soit $y \in x$. De plus, tous les éléments d'un ordinal sont des ordinaux. On notera parfois $x < y$ à la place de $x \in y$ et $x \leq y$ pour $x \subset y$. L'opération successeur se note s et, pour tout ordinal α , vérifie $s(\alpha) = \alpha \cup \{\alpha\}$. En particulier, les ordinaux finis sont identifiés avec les naturels. Zéro est l'ensemble vide $\{\}$ et, en appliquant l'opération s de façon récursive, on définit $n + 1 = n \cup \{n\} = \{0, 1, \dots, n\}$. L'ensemble de tous les ordinaux fini est ω_0 , noté aussi ω . C'est le plus petit ordinal infini, et il coïncide avec N . Le premier ordinal indénombrable est ω_1 . Le deuxième ordinal indénombrable à ne pouvoir être écrit comme une opération sur des ordinaux inférieurs est ω_2 . En poursuivant cette même logique, l'on définit $\omega_3, \omega_4, \dots, \omega_{\omega_0}$, etc.

Pour tout ordinal α , l'on note $\cup(\alpha)$ l'union de tous les éléments de α . Soit α possède un élément maximal $\beta = \cup(\alpha)$ et alors on a $s(\beta) = \alpha$. Soit il n'en possède pas et alors $\cup(\alpha) = \alpha$. Dans le premier cas, l'on dit de α qu'il est un ordinal non-limite. Dans le deuxième cas, c'est un ordinal limite et l'on note $\alpha \in \text{Lim}$. Les ordinaux de la forme ω_α sont tous des ordinaux limites.

L'addition "+" et la multiplication "." pour les ordinaux sont définis par un ensemble d'axiomes. Étant donnés deux ordinaux α et β :

$$\begin{aligned} \alpha + 0 &= \alpha \\ \alpha + s(\beta) &= s(\alpha + \beta) \\ \alpha + \beta &= \cup \alpha + \lambda : \lambda < \beta \text{ si } \beta \in \text{Lim} \\ \alpha \cdot 0 &= 0 \\ \alpha s(\beta) &= s(\alpha \beta) \\ \alpha \beta &= \cup \alpha \lambda : \lambda < \beta \text{ si } \beta \in \text{Lim} \end{aligned}$$

L'opération $s(\alpha)$ pourra aussi s'écrire $\alpha + 1$. Notons que l'addition et la multiplication sont associatives, mais pas commutatives. Pour tout naturel

n , on a en particulier $n + \omega_0 = \omega_0 < \omega_0 + n$. De même, $2\omega_0 = \omega_0 < \omega_0 + \omega_0 = \omega_0 2$.

Dans la mesure où l'on accepte l'axiome du choix, l'on peut définir le cardinal α d'un ensemble E comme le plus petit ordinal de von Neumann qui lui soit équipotent, i.e. l'on identifie le cardinal à son ordinal initial. Les cardinaux finis s'apparentent ainsi aux naturels. Zero est le cardinal de l'ensemble vide, $n > 0$ est cardinal de $\{0, \dots, n-1\}$. La cardinalité de l'ensemble des naturels N est dénoté $\aleph_0$. Cela dit, $\aleph_0$ est aussi la cardinalité de tous les ordinaux infinis dénombrables, puisque ces ordinaux sont équipotents à N . $\aleph_0$ peut donc être identifié indifféremment à $\omega_0, \omega_0^2, \omega_0^{\omega_0}$, etc. L'arithmétique sur les ordinaux n'est pas la même que sur les cardinaux : bien que $\omega_0^2 > \omega_0$, on a $\aleph_0^2 = \aleph_0$. En ce sens, un cardinal est une mesure de la "taille" d'un ensemble, sans tenir compte de son ordre.

Le cardinal infini qui est directement supérieur à $\aleph_0$ est noté $\aleph_0^+$, ou encore $\aleph_1$. Ce cardinal peut être identifié à ω_1 , ainsi qu'à tous les ordinaux résultant d'un nombre fini d'opérations sur ω_1 . Notons que si l'on admettait l'hypothèse du continu, l'on aurait $\aleph_1 = 2\aleph_0$, la cardinalité de l'ensemble des réels R . On a ensuite $\aleph_1^+ = \aleph_2$, puis $\aleph_2^+ = \aleph_3$, etc. De façon générale, tout cardinal infini s'écrit $\aleph_\theta$, où θ est un nombre ordinal.

L'on dit du cardinal $\aleph$ d'un ensemble E qu'il est régulier lorsqu'il est égal à sa propre cofinalité, c'est-à-dire lorsqu'il est égal à la cardinalité du plus petit ensemble E' tel que pour tout $a \in E$, il existe $b \in E'$ avec $a \leq b$. Sous l'axiome du choix, l'on peut dire de façon équivalente que α est régulier si pour tout ensemble d'indices I et pour toute suite d'ensembles $S_i (i \in I)$ tels que $\text{card}(I) \leq \alpha$ et $\text{card}(S_i) < \text{card}(I)$, l'on a $\text{card}(\cup S_i : i \in I) < \alpha$. C'est cette deuxième définition qui sera utilisée le plus souvent. Notons que $\aleph_0$, ainsi que tous les cardinaux successeurs sont réguliers. En revanche, $\aleph_\omega$ est singulier.

Un cardinal $\aleph_\theta$ est dit limite si θ est un ordinal limite. Un cardinal $\aleph_\theta$ est dit faiblement inaccessible s'il est limite et régulier. Il est dit fortement inaccessible s'il vérifie en plus la condition $\lambda < \aleph \rightarrow 2^\lambda < \aleph$. Tous les cardinaux inaccessibles sont de grands cardinaux, supérieurs à $\aleph_\omega$, et construits à l'aide des cardinaux plus petits. Sous les axiomes ZFC, il n'est pas possible de démontrer l'existence d'un cardinal inaccessible. En effet, la théorie ZFC est supposée cohérente. Or ajouter la proposition "Il n'existe pas de cardinal inaccessible" à ZFC donne une théorie également cohérente. Par suite, le problème est indécidable. Il faut noter que sous l'hypothèse du continu, les notions d'inaccessibilité faible et d'inaccessibilité forte se confondent.

Annexe 6 : La Théorie de l'information de Shannon

L'information est la
détermination de l'incertitude.

Claude Shannon

L'information peut être entendue en bien de sens : du point de vue sémantique, elle est le contenu intuitif et cognitif d'un énoncé, sa signification, et peut être vue comme ce que cet énoncé enseigne sur le monde. A ce titre, des lapalissades du type "S'il n'était pas mort, il serait encore en vie" n'enseignent rien qui ne soit déjà connu *a priori* et ne constituent donc pas des informations. De même, des redondances comme "Le cheval blanc de Napoléon est blanc" sont plus pauvres en information que des propositions innovantes pour la science. Enfin, la quantité d'information gagnée par le fait de savoir qu'un événement donné s'est réalisé ou doit se réaliser est d'autant plus grande que la probabilité de cet événement est faible. Si l'on apprend, par exemple, que conformément aux statistiques il a plu sur toute la côte de Bretagne, l'information gagnée est moindre que si l'on constate, contre toute attente, qu'il s'est produit une canicule.

D'un point de vue syntaxique en revanche, l'information est la quantité de symboles logiques ou linguistiques requis pour transcrire un énoncé dans un langage donné. Au plus ce nombre est important, au plus l'information est dense. En ce sens, les tautologies sont des informations à part entière, d'autant plus importantes qu'elles exigent plus de caractères pour être exprimées ; inversement, des propositions très riches de sens, comme " $e^{i\pi} = -1$ ", peuvent être formulées de façon lapidaires, et être donc très pauvres en information du point de vue syntaxique.

A première vue, ces deux approches de ce que constitue une information paraissent difficiles à concilier, et sont à bien des égards exclusives l'une de l'autre. Il est pourtant légitime de se poser les deux questions suivantes : (1) Est-il possible de les formaliser l'une et l'autre dans un langage mathématique à l'aide de formules quantifiables commodes du point de vue de la manipulation, et suffisamment riches pour capturer l'intuition ? (2) Le cas

échéant, ces deux formules distinctes peuvent-elles être généralisées en une seule et même formule plus ample, au vu de laquelle les deux définitions précédentes ne seraient que des cas particuliers ?

La théorie statistique de l'information a répondu par l'affirmative à chacune de ces deux interrogations et elle a exhibé, au terme de sa longue maturation, un concept de quantité d'information connu sous le nom d'entropie. Introduite de façon inaugurale par Shannon dans son article intitulé *A Mathematical Theory of Communication* (1948), cette théorie a le mérite de synthétiser les approches axiomatiques existantes et d'en préciser les principaux domaines d'application.

Dans ce qui suit, les fondements de la théorie de l'information sont récapitulés, et une démarche axiomatique est présentée. Ensuite, les principales mesures de l'information et leurs propriétés sont énoncées et, tour à tour, démontrées.

Point de vue historique et mathématique

Les approches de Hartley et de Wiener

Hartley propose en 1928 une définition de la quantité d'information basée sur l'espace mémoire, ou le nombre de bits, qui sont nécessaires pour encoder un ensemble quelconque d'éléments. En présence de 2^k éléments par exemple, l'on sait que k bits suffiront. Par extension, soit un ensemble $\mathcal{X}$ de taille quelconque. L'on écrira

$$H_{\text{Hartley}}(\mathcal{X}) = \log_2(|\mathcal{X}|).$$

Si $\mathcal{X}$ est un alphabet, il peut être utilisé pour écrire un mot x . La quantité d'information contenue dans x correspondra au nombre de bits requis pour encoder les n lettres de ce mot :

$$H_{\text{Hartley}}(x) = n \log_2(|\mathcal{X}|).$$

Selon ce point de vue, la quantité d'information ne dépend que de la taille de l'alphabet, sans tenir compte, ni de la signification des mots, ni de la fréquence des lettres de cet alphabet. Or un codage plus intelligent permettrait aisément d'exploiter les récurrences de certaines lettres de façon à réduire l'espace mémoire requis, en assignant des codes de longueur plus courte aux mots les plus fréquents, et des longueur plus grande aux moins fréquents.

Cette observation suggère à Wiener une autre définition de la quantité d'information, basée non plus sur la taille de l'alphabet, mais sur la fréquence des lettres qu'il contient. L'intuition sous-jacente est la suivante. Supposons que le lecteur souhaite décoder un message. Chaque lettre qu'il parvient à déchiffrer lui apporte de l'information sur le message, qui est d'autant plus

précieuse que cette lettre est rare dans la langue française. Une mesure H de la quantité d'information acquise par la lecture d'une lettre a est donc fonction de la fréquence p de cette lettre dans l'alphabet. En particulier, $H(p)$ doit être nulle lorsque $p = 1$ et tendre vers l'infini lorsque p tend vers 0. Entre ces deux valeurs extrêmes, $H(p)$ évolue de façon monotone. En 1948, Wiener construisit une fonction satisfaisant ces propriétés :

$$H_{\text{Wiener}}(p) = -\log_2(p).$$

De plus, il proposa un ensemble d'axiomes sous lesquels cette fonction est univoquement définie. Au nombre de trois, ils s'énoncent de la façon suivante : soit un événement quelconque de probabilité p et un autre de probabilité q , H_{Wiener} est la fonction telle que :

- (i) **Positivité** $H_{\text{Wiener}}(p) \geq 0$,
- (ii) **Additivité** $H_{\text{Wiener}}(pq) = H_{\text{Wiener}}(p) + H_{\text{Wiener}}(q)$,
- (iii) **Normalisation** $H_{\text{Wiener}}(\frac{1}{2}) = 1$.

Le premier axiome se justifie en ceci que toute lecture apporte de l'information. Dans le pire des cas, cette information n'est pas utile, mais elle ne peut pas réduire la connaissance du lecteur. Le troisième axiome est posé à titre de normalisation, de façon à garantir l'unicité. L'axiome (ii) quant à lui est le plus chargé en terme de signification et mérite une discussion. Supposons que le lecteur soit amené à lire successivement deux lettres a et b appartenant à un alphabet $\mathcal{X}$, de fréquences respectives p et q . Appelons A l'événement consistant à lire la lettre a et B l'événement consistant à lire la lettre b . Supposons en outre que la lecture de ces deux lettres constituent des événements statistiquement indépendants l'un de l'autre. Sous cette condition, la probabilité de l'événement $A \cap B$, ou autrement dit la probabilité de lire les deux lettres l'une à la suite de l'autre (quel que soit l'ordre), est pq . Or la quantité d'information apportée par $A \cap B$ est la somme de l'information apportée par A et de l'information apportée par B . On peut donc écrire $H_{\text{Wiener}}(pq) = H_{\text{Wiener}}(p) + H_{\text{Wiener}}(q)$.

Il est facile de vérifier que, sous ces axiomes, H_{Wiener} vérifie bien la décroissance monotone. Soient en effet deux probabilités p_1 et p_2 telles que $0 \leq p_1 \leq p_2 \leq 1$ et écrivons sans perte de généralité $p_1 = qp_2$, avec $0 \leq q \leq 1$. L'on a :

$$\begin{aligned} H_{\text{Wiener}}(p_1) &= H_{\text{Wiener}}(qp_2) \\ &= H_{\text{Wiener}}(q) + H_{\text{Wiener}}(p_2) && \text{(par l'axiome ii)} \\ &\geq H_{\text{Wiener}}(p_2) && \text{(par l'axiome i)} \end{aligned}$$

Notons que la base logarithmique utilisée dans la notion d'entropie peut être prise égale à $a \neq 2$. Il suffit pour se faire de changer l'axiome de norma-

lité, et d'écrire $H_{\text{Wiener}}(1/a) = 1$ au lieu de $H_{\text{Wiener}}(1/2) = 1$. Dans le choix du logarithme népérien ($a = e$), l'unité est le nat.

L'approche de Shannon

La question que nous adressons maintenant consiste à savoir quelle quantité d'information gagne en moyenne le lecteur lorsqu'il prend connaissance de la première lettre d'un mot aléatoire? L'on suppose que la lettre lui est transmise par un émetteur E au moyen d'un canal parfait (i.e. non bruité) et qu'il n'y a donc pas d'erreur de lecture. Dans ce cas de figure, la lettre découverte par le lecteur est représentée par une variable aléatoire X à valeur dans $\mathcal{X}$. Supposons que $m = |\mathcal{X}|$ et notons p_i la fréquence de la $i^{\text{ème}}$ lettre de l'alphabet $\mathcal{X}$ (avec $i = 1, \dots, m$). La probabilité de lire un a dépend de la fréquence p_a de la lettre a et apportera une information égale à $H_{\text{Wiener}}(p_a)$, la probabilité de lire un b dépend de la fréquence p_b de la lettre b et apportera une information égale à $H_{\text{Wiener}}(p_b)$, etc... La quantité d'information liée à la lecture de cette variable aléatoire est donc donnée par l'espérance selon X , notée E_x , de l'entropie de Wiener :

$$\begin{aligned} H(X) &= E_X[H_{\text{Wiener}}(p_i)] \\ &= - \sum_{i=1}^m p_i \log(p_i). \end{aligned}$$

Cette quantité est appelée l'entropie de Shannon. Elle est d'autant plus importante que les probabilités associées à chacune des lettres de l'alphabet sont comparables les unes aux autres, et atteint un maximum dans le cas de l'équiprobabilité. Si tous les symboles ont la même fréquence en effet, le lecteur ne possède aucun indice pour deviner de façon apriorique quelle est l'état de X et c'est pourquoi il acquière une plus grande quantité d'information en découvrant cet état. Illustrons ceci à l'aide d'un exemple.

Exemple 9. *Considérons la variable aléatoire X suivante :*

$$X = \begin{cases} a & \text{avec une probabilité } p \\ b & \text{avec une probabilité } 1 - p \end{cases}.$$

Si $p = 1$ ou $p = 0$, l'on a toujours $X = a$ ou $X = b$ (respectivement). Dans ce cas, il va de soi que la lecture du symbole n'apporte aucune information que l'on ne connaisse déjà. En revanche, si l'on a $0 < p < 1$, l'on peut déduire que $X = a$ avec une probabilité p , mais il n'est pas possible de le savoir avec certitude. L'information fournie par l'émetteur permet de lever cette incertitude et apporte de l'information. L'information est ainsi liée à

l'incertitude. Au plus cette incertitude est grande, au plus l'information est intéressante. La courbe $H(p) = -p \log(p) - (1-p) \log((1-p))$ est reprise sur le graphique ci-dessous.

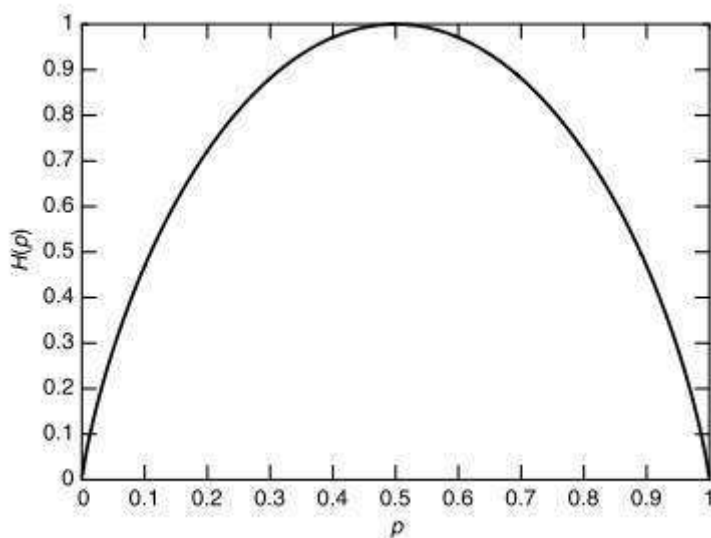


FIGURE 16.4 – $H(p)$ en fonction de p .

La définition d'entropie peut aussi être étendue à des séquences d'événements dont la somme des probabilités est inférieure à 1. Soient $p_1, \dots, p_m \in [0, 1]$ tels que $\sum_{i=1}^m p_i \leq 1$. L'on écrira H_m la fonction entropie à m arguments définie de la façon suivante :

$$H_m(p_1, \dots, p_m) = - \sum_{i=1}^m p_i \log(p_i).$$

Remarquons que cette écriture généralise l'entropie de Hartley et de Wiener. Dans le cas où $p_1 = \dots = p_m = 1/m$, en effet, l'on a $pH(X) = \sum_{i=1}^m 1/m \log(m) = \log(m) = H_{\text{Hartley}}(\mathcal{X})$. Dans le cas où $m = 1$, l'on a $H_1(p_1) = -\log(p_1) = H_{\text{Wiener}}$.

Système axiomatique

Sous certains axiomes, l'entropie de Shannon peut être dérivée de façon univoque. Le choix des axiomes peut varier dans la littérature (cf. [?], [?]), mais tous conduisent à la même définition de $H(X)$. Les axiomes que nous présentons ici sont ceux qui ont été optés par Shannon lui-même.

— Axiomes de Shannon —

- **Normalisation** : $H_2(\frac{1}{2}, \frac{1}{2}) = 1$.
- **Continuité** : Pour tout $p \in [0, 1]$, $H_2(p, 1 - p)$ est une fonction continue de p .
- **Groupement** : Pour tout nombre naturel m et pour toute suite de réels $p_1, \dots, p_m$ telle que $\sum_{i=1}^m p_i \leq 1$:

$$H_m(p_1, \dots, p_m) = H_{m-1}(p_1 + p_2, p_3, \dots, p_m) + (p_1 + p_2)H_2(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}).$$

Il est possible de montrer que l'unique fonction H_m satisfaisant ces trois axiomes s'écrit :

$$H_m = - \sum_{i=1}^m p_i \log(p_i)$$

En raison de la longueur et de la complexité de cette preuve, nous prions le lecteur de se référer aux annexes. Nous nous contenterons de donner ici les étapes clef de la démarche et de les commenter.

L'idée fondamentale consiste à caractériser la fonction $H_m(\frac{1}{m}, \dots, \frac{1}{m})$, qui correspond à l'entropie d'une variable aléatoire uniformément distribuée, et à montrer à partir des axiomes qu'elle est égale à $\log(m)$. Cette démarche est comparable à celle de Wiener, si ce n'est que l'ensemble d'axiomes de départ n'est pas le même. Pour y parvenir, il est nécessaire de démontrer et puis d'utiliser le lemme suivant :

Théorème 12. *Si la fonction $f(m)$ satisfait :*

- $f(mn) = f(m) + f(n)$
- $\lim_{n \rightarrow \infty} (f(n+1) - f(n)) = 0$.
- $f(2) = 1$

alors $f(m) = \log_2(m)$.

Une fois cela établi, il devient possible de montrer que pour tout rationnel p , $H_2(p, 1 - p) = -p \log(p) - (1 - p) \log(1 - p)$. En utilisant la continuité, ce résultat peut être étendu au irrationnel. Il est finalement généralisé à tous les $m \geq 2$ par une procédure d'induction utilisant l'axiome de groupement (sous une forme généralisée).

Notons que si l'on ajoute l'axiome de monotonie, la preuve du lemme ci-dessus s'en trouve grandement simplifiée. Cet axiome énonce que pour tout $m \in \mathbb{Z}^+$, $f(m) = H_m(\frac{1}{m}, \dots, \frac{1}{m})$ est une fonction de m croissante et monotone.

Il existe de nombreuses alternatives à la démarche axiomatique proposée ci-dessus. A peu près toutes les propriétés de l'entropie ont été utilisées pour

démontrer axiomatiquement la formule de l'entropie. La difficulté est de les choisir en nombre le plus restreint possible, et de les choisir de telle sorte qu'ils soient intuitivement parlant.

Les Mesures de Shannon

Dans la section qui suit, un récapitulatif des différentes grandeurs entropiques va être donné : entropie jointe, entropie conditionnelle, divergence de Kulback-Leibler, information mutuelle, etc... Nous récapitulons d'abord l'ensemble des définitions, après quoi nous prouvons les principales inégalités faisant intervenir ces mesures. Les définitions et les preuves sont établies sous la forme multivariée.

Entropie jointes et conditionnelles

Pour tout $n \geq 2$, considérons les variables aléatoires $X_1, \dots, X_n$ discrètes à valeurs dans $\mathcal{X}_1, \dots, \mathcal{X}_n$ (respectivement). A ces n variables aléatoires sont associées $k = 2^n - 1$ fonctions de probabilité jointes. Posons $\mathcal{N} = \{1, \dots, n\}$. Pour tout $\alpha = (i_1, \dots, i_m) \subseteq \mathcal{N}$, l'on dénote par p_α la fonction de probabilité jointe correspondant aux variables $X_{i_1}, \dots, X_{i_m}$. Suivant la même logique, l'on note aussi $X_\alpha = \{X_i : i \in \alpha\}$, $\mathcal{X}_\alpha = \{\mathcal{X}_i : i \in \alpha\}$. L'on peut étendre la notion d'entropie à un ensemble de plus de deux variables aléatoires.

Entropie jointe

Soit $\alpha \subseteq \mathcal{N}$. L'entropie jointe des variables X_α s'écrit

$$H(X_\alpha) = H(\alpha) = - \sum_{x \in \mathcal{X}_\alpha} p_\alpha(x) \log(p_\alpha(x)).$$

Par convention, l'on suppose que $H(\emptyset) = 0$.

De la même façon que l'on peut parler de probabilité conditionnelle, il est possible de parler d'entropie conditionnelle. Cette entropie est définie en substituant les probabilités conditionnelles aux probabilités intervenant dans la définition de l'entropie.

Entropie conditionnelle

Soient $\alpha, \gamma \subseteq \mathcal{N}$. L'entropie conditionnelle des variables X_α par rapport aux variables X_γ s'écrit :

$$H(X_\alpha | X_\gamma) = H(\alpha | \gamma) = - \sum_{x \in \mathcal{X}_\alpha} \sum_{y \in \mathcal{X}_\gamma} p_\alpha(x|y) \log(p_\alpha(x|y)).$$

Par convention, l'on écrira $H(\alpha | \beta) = H(\alpha)$ lorsque $\beta = \emptyset$.

Entropie relative

L'entropie relative, encore appelée divergence de Kullback-Leibler, définit une pseudo-distance dans l'espace des distributions statistiques. Elle vise à quantifier la différence qui existe entre deux densités de probabilités p et q . En particulier, elle est nulle lorsque p et q sont identiques sur $\text{supp}(p)$, le support compact de p . Au sinon, elle est toujours positive. L'entropie relative ne nous apprend rien sur p et q en dehors de $\text{supp}(p)$, ce qui en constitue la principale limitation.

Entropie relative

Soient deux densités de probabilité p et q définies sur un même domaine $\mathcal{X}$. L'entropie relative (ou divergence de Kullback-Leibler) entre p et q , notée $D(p||q)$, est définie comme suit :

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) f(x),$$

où f est le prolongement continu de $\log \frac{p(x)}{q(x)}$

$$f(x) = \begin{cases} 0 & \text{si } p(x) = 0, \\ \infty & \text{si } p(x) \neq 0 \text{ et } q(x) = 0, \\ \log\left(\frac{p(x)}{q(x)}\right) & \text{sinon.} \end{cases}$$

Notons que s'il existe un $x \in \mathcal{X}_\alpha$ tel que $q(x) = 0$ et $p(x) \neq 0$, l'entropie relative n'est pas bornée.

L'entropie relative vérifie deux des trois axiomes d'une métrique. C'est une pseudo-distance, et non une distance au sens strict, car l'inégalité de Minkowski n'est pas satisfaite. De plus, $D(p||q)$ n'est pas symétrique en général. En effet, s'il existe un $x \in \mathcal{X}$ tel que $p(x) = 0$ et $q \neq 0$, on a $D(p||q) = 0$ tandis que $D(q||p)$ n'est pas bornée.

$$D(p||q) = 0 \Leftrightarrow p(x) = q(x) \quad \forall x \in \text{supp}(p), \quad (16.1)$$

$$D(p||q) \geq 0. \quad (16.2)$$

La notation $\text{supp}(p)$ fait référence au support compact de p . La preuve de ces deux propriétés est fournie en annexes.

Information mutuelle

Une autre grandeur d'importance en théorie de l'information est l'information mutuelle. Cette quantité, définie à base de l'entropie, mesure l'information qui est détenue sur un ensemble de variables aléatoires lorsque l'on connaît l'état d'un autre ensemble de variables aléatoires, statiquement dépendants. C'est, en d'autres termes, une mesure de la perte d'incertitude

acquise sur un ensemble de variables aléatoires en raison de la connaissance d'un autre ensemble de variables aléatoires.

— Information mutuelle —

Soient deux ensembles $\alpha, \beta \subseteq \mathcal{N}$. Notons $p_{\alpha \cup \beta}$ la fonction de probabilité correspondante à l'union des ensembles de variables X_α et X_β . *L'information mutuelle* est notée $I(X_\alpha; X_\beta)$ et vaut :

$$I(\alpha; \beta) = D(p_{\alpha \cup \beta} \| p_\alpha p_\beta).$$

Pour tout ensemble $\gamma \subseteq \mathcal{N}$, l'on peut définir *L'information mutuelle conditionnelle* :

$$I(\alpha; \beta | \gamma) = D(p_{\alpha \cup \beta | \gamma} \| p_{\alpha | \gamma} p_{\beta | \gamma}).$$

Par convention, l'on écrira $I(\alpha; \beta | \gamma) = I(\alpha; \beta)$ lorsque $\gamma = \emptyset$.

Si l'ensemble de variables X_α est statistiquement indépendant de l'ensemble de variables X_β , alors pour tout $(x, y) \in \mathcal{X}_\alpha \cup \mathcal{X}_\beta$, $p_{\alpha \cup \beta}(x, y) = p_\alpha(x)p_\beta(y)$ et $D(p_{\alpha \cup \beta} \| p_\alpha p_\beta)$ est nulle. La réciproque n'est pas vraie. Supposons que $p_{\alpha \cup \beta} = p_\alpha p_\beta$ partout sur le support compact de $p_{\alpha \cup \beta}$, mais qu'il existe un point (x, y) n'appartenant pas à celui-ci et tel que $p_\alpha p_\beta$ soit non-nul. Dans ce cas, on a $D(p_{\alpha \cup \beta} \| p_\alpha p_\beta) = 0$. Pourtant, X_α et X_β ne sont pas statistiquement indépendants. A cette restriction près, l'information mutuelle peut être vue comme une mesure de la dépendance statistique. Cette grandeur n'est jamais négative (car l'entropie relative n'est jamais négative).

Annexes 7 : Essai - La Conscience est-elle morte ?

Cette thèse pose les fondations d'une théorie méréologique de la conscience, mais laisse un certain nombre de questions en suspens. Pour commencer, il est légitime de se demander comment les notions du Je, du Moi et du Soi, que je n'ai fait qu'esquisser de façon grossière, s'articulent avec les trois niveaux cognitifs que sont la Sensation, la Perception et la Raison ? Par ailleurs, il semble essentiel de justifier l'introduction du principe de différenciation, dont j'ai dit qu'il était au cœur du mécanisme de la conscience.

Pour élucider ces interrogations, et donner à mon propos plus de consistance, j'ai écrit un essai littéraire de 300 pages intitulé : « La Conscience est-elle morte ? » Ce texte, dans sa version préliminaire, a reçu le premier prix du concours Art et Lettres de France en 2022, dans la section essai. En dépit de cette récompense, je n'ai cessé de le retravailler, jusqu'à lui donner la forme que je vous présente aujourd'hui. En raison de son style poétique, du ton personnel et polémique que j'adopte à l'occasion, ainsi que des libertés que je m'autorise dans le développement des idées, j'ai jugé préférable de ne pas l'adjoindre au corps de cette thèse. Il ne s'agit en effet pas d'un discours académique classique, qui discute des théories existantes en vue de se positionner au sein d'un débat d'experts. Il s'agit plutôt d'une prise de position personnelle, humaniste et universaliste, face à une question touchant à la condition même de l'existence humaine : que suis-je en tant qu'être conscient ? qu'est-ce qui donne encore de la valeur au monde d'aujourd'hui ?

Le lecteur trouvera ci-joint une courte présentation de l'ouvrage. Pour une meilleure compréhension de son articulation, je l'invite également à se référer à la table des matières du texte proprement dit.

Présentation de l'essai

Lors d'un voyage dans le nord-ouest de la Thaïlande, il m'est arrivé de faire une étrange rencontre. Il faisait chaud et humide, le soleil tombait à la verticale, nimbant les chemins poussiéreux d'un halo de lumière rouge sang. De la sueur me perlait le long du corps et les moustiques s'en donnaient

à cœur joie. C'était quelques mois avant le début de mon doctorat. J'étais loin de me douter que les peines encourues, en cette jungle inhospitalière, n'étaient que le prélude d'un long périple dans le monde des idées.

En dépit de la chaleur accablante, il ne s'écoula pas longtemps avant que j'arrive aux abords d'un monastère. Au moment d'entrer dans la cour intérieure, j'ai entendu des voix non loin de moi. Il se trouvait là un jeune homme qui parlait avec faconde, agitant les mains à chaque mot — à moins que ce ne fût à cause des moustiques ? À ses côtés, un moine, plus âgé, écoutait silencieusement. En m'approchant plus près d'eux, j'ai constaté qu'ils discutaient d'un sujet proche de ma thèse : « En raison des connaissances accumulées sur le cerveau, affirmait le jeune homme d'un geste passionné, il est aujourd'hui permis de penser que la conscience, telle que nous la vivons, pourrait être une illusion. Vous savez ce qui m'a le plus saisi en commençant mes études en neurologie ? Que toutes mes sensations ne sont rien que des courants électriques. » Face à lui, le moine se contentait de sourire. Il posait çà et là des questions, sans donner son point de vue. À la fin de la conversation, il lui a proposé d'aller manger. Mais au moment de se lever, il s'est ravisé : « Supporteras-tu, a-t-il demandé, l'illusion de la nourriture ? »

Cette réplique m'a fait sourire, car je la trouve assez illustrative d'une tension existant, depuis le début de l'ère moderne, entre la science et l'expérience personnelle. Du point de vue purement intime et introspectif, la seule chose indubitable pour un moine bouddhiste est d'être, à un instant donné, conscience de quelque chose plutôt que rien. Le monde tangible qui se déploie sous le tâtonnement de ses doigts, de ses yeux, de sa langue, n'existe que dans la conscience qu'il en a. L'idée même d'être un sujet percevant le monde, un corps au sein de la matière, une personnalité pensante dotée de son histoire, de sa mémoire, de son caractère immuable, semble pouvoir être remise en cause. Seul existe l'instant dans sa plénitude.

Du point de vue de l'expérience scientifique en revanche, il semblerait que la certitude première soit l'existence d'un monde matériel, fait de hadrons, d'ondes et d'énergie, au sein desquels il est difficile d'expliquer comment émerge le phénomène de la conscience proprement dit. Lorsque je regarde le ciel sempiternel, que je me représente, un à un, l'infinité des mondes piqués sur la grande toile de la Voie lactée, comme tout autant d'étoiles en perdition dans le vide glacé de l'abîme, je ne puis m'empêcher de ressentir un curieux vague à l'âme. L'Univers me semble effroyablement plus réel que la conscience que j'en ai. Il est peuplé de sons, de couleurs, de fragrances enivrantes, qui ne sont rien d'autre que des particules quantiques en déplacement stochastique. Ce que je prends pour un authentique phénomène de conscience (la contemplation du ciel) n'est qu'un peu de bruit et de fureur (des réactions neurologiques dans le cerveau).

Tout au long de cet essai, je vais m'efforcer d'élaguer l'arbre de la métaphysique occidentale — ce tronc calciné par des années de luttes philosophiques — afin de le soulager de l'emprise de ces deux préjugés. Le premier

préjugé consiste à dire (1) que la conscience est dans le cerveau. Le second préjugé, au contraire, affirme (2) que le monde est tout entier dans la conscience qu'on en a. Ces deux suppositions sont aussi vieilles que le matérialisme et que l'idéalisme. Elles s'opposent depuis des temps immémoriaux sous la plume amère des philosophes en robe de chambre, mais peut-être n'avaient-elles jamais encore été si coriaces ? Et pour cause : une seule et même idée est nichée derrière ces deux opinions. Cette idée, comme je m'efforcerai de le montrer dans cet essai, c'est que la conscience soit quelque chose d'individuel.

On admet comme une évidence que la conscience est un fait personnel ; toutes les notions, les sensations et perceptions qui nous traversent nous appartiennent en propre. L'idée que notre esprit puisse contenir des vécus qui ne nous appartiennent pas nous semblerait étrange. C'était pourtant la croyance du peuple des Ashuars, en Amazonie, qui voyait le rêve comme un espace où leurs pensées se mêlent à celles des plantes et des animaux. Il était également admis, dans le giron de la religion chrétienne, qu'un croyant puisse entendre la parole de Dieu en lui. Il n'était dès lors pas si rare de représenter la conscience comme un tout qui dépassait notre esprit individuel ; la vie mentale de l'être humain était imbriquée dans celle de Dieu (christianisme), de Brahma (hindouisme), de l'Esprit (Spinoza), de l'Absolu (Hegel), de l'Un (Plotin) ou du Noûs (Anaxagore).

Il en va tout différemment aujourd'hui. Celui qui rapporte avoir des transmissions de pensées, sentir en lui la présence d'un mort ou entrer en communication avec une conscience supérieure (la nature, dieu, un esprit...) est soupçonné de présenter une malformation neurologique. Depuis que l'on tient la conscience pour un fait individuel, qui existe de façon monadique, indépendamment de toute autre conscience, seules deux possibilités métaphysiques s'ouvrent encore à la philosophie : soit la conscience est nichée quelque part dans le corps de l'individu (c'est la posture du matérialisme réductionniste), soit le monde entier est compris dans la conscience de cet individu (c'est la posture de l'idéalisme solipsiste). Ce tiers exclu méthodologique nous empêche de penser que la conscience puisse dépasser les frontières de notre corps (posture du cognitivisme étendu) ou qu'elle puisse dépasser les frontières de notre Moi (posture de l'idéalisme non solipsiste).

Dans cet essai, je suggérerai que ces deux attitudes, celle qui fait de la conscience une propriété du cerveau (matérialisme réductionniste) et celle qui fait du monde matériel une propriété de la conscience (idéalisme solipsisme), conduisent à des contradictions. Elles échouent en effet à articuler, dans l'étude de l'esprit humain, deux notions bien différentes :

- **La Conscience subjective** : D'une part, la conscience désigne un état. C'est-à-dire, le fait subjectif d'être conscient, *ici et maintenant*, de quelque chose plutôt que de rien. Il ne s'agit toutefois pas du fait d'être conscient d'une chose en particulier : la table qui est devant

moi, le vent dans mes cheveux, etc. Il s'agit bien plutôt de la faculté d'être conscient de potentiellement tout et n'importe quoi. Tant que nous la possédons, cette faculté est ce qui nous fait dire que nous sommes en vie. C'est pourquoi je la désignerai aussi sous le nom de *conscience vive*.

- **La Conscience objective** : D'autre part, la conscience désigne également un objet du monde doté de propriétés physiques ou psychologiques : l'esprit, le Moi, l'âme, le cerveau, etc. On impute à cet objet, aussi concret que la représentation d'une table devant moi ou que la sensation du vent dans mes cheveux, la cause de toute expérience consciente possible. Cette conscience objective n'est donc pas le processus mental dans sa totalité. Ce n'est ni un phénomène organique, ni une aptitude. C'est une *conscience morte*, qui attend, comme tous les autres objets observables, d'être disséquée par le scalpel de la méthode scientifique.

Dans le matérialisme réductionniste, la conscience objective prend la forme d'un cerveau doté de neurones, de synapses, de matière grise, dont l'activité exubérante produirait la pensée. Dans l'idéalisme solipsiste, la conscience objective est assimilée à une sorte de réceptacle passif, un Moi qui serait à l'origine du mental — et auquel toutes les expériences vécues pourraient être attribuées. Dans un cas comme dans l'autre, on procède à un tour de passe-passe ingénieux, consistant à donner une forme objective, spatiale, temporelle, à notre conscience subjective — qui n'est qu'un simple *ici et maintenant*.

Cet escamotage conceptuel se heurte à un paradoxe logique, dont il est possible de donner une analogie instructive. Supposons qu'un individu ait grandi toute sa vie dans une pièce dépourvue de toute surface réfléchissante. Il est capable de voir le monde autour de lui : les murs, les tables, les chaises, etc. Il peut également contempler ses mains, son corps et — avec un peu de chance — le bout de son nez rougi par une trop grande quantité d'alcool, mais il n'a jamais vu son propre visage, ni la couleur de ses yeux. Comment un tel homme peut-il expliquer le phénomène de la vision ?

Notre individu hypothétique comprendrait certes que la vision, en tant qu'expérience quotidienne, doit avoir une origine en lui. Il pourrait même affirmer que cette origine se trouve quelque part sur son visage, mais, bien qu'il puisse fermer ses paupières, palper ses yeux du bout des doigts, il ne serait pas capable de voir ses yeux de ses propres yeux. Autrement dit, la vision serait pour lui une évidence sensitive, mais pas une évidence d'ordre visuel. Il aurait conscience de son propre œil comme de quelque chose d'abstrait, qu'il peut toucher avec ses mains, qu'il peut sentir en lui, mais qu'il ne peut pas voir.

Que devient cette analogie lorsqu'on la rapporte, non pas à l'œil de la

vision, mais à la conscience ? Au même titre que l'œil ne tombe pas lui-même sous le sens de la vue, il se pourrait que la conscience ne soit pas, en tant qu'expérience, appréhendable par les ressources de notre conscience. Chaque fois que je me représente moi-même en tant que conscience, que je me dis « Je suis un tel », « Je suis celui qui pense », « Je suis cogito », « Je suis le sujet de toute expérience » ou « je suis un cerveau dans le monde » ; chaque fois que je me livre à mon auto-contemplation, j'ai sous les yeux une image, une émotion, une sensation de moi-même. Cette dernière est un objet de ma pensée comme tant d'autres, et ne se distingue pas en cela de la représentation d'une table, d'un livre ou d'un singe. Cette image de moi n'est pas ma conscience objective en tant que telle, mais une simple représentation subjective contenue dans mon expérience consciente.

Cet apparent paradoxe est peut-être la raison fondamentale pour laquelle le monde conscient semble peuplé d'objets qui sont en moi mais qui ne sont pas *moi* ? La grande idée de mon essai consistera à défendre que la conscience, contrairement aux objets usuels, ne peut pas être représentée. Elle peut soit être *vécue* subjectivement comme un « Je » (je suis conscience), soit *conçue* abstraitement et objectivement comme un « Il » (il y a de la conscience). Dans le premier cas, la conscience est alors un je-ne-sais-quoi dont on ne peut rien dire. Elle se manifeste, mais dès l'instant où elle apparaît, elle n'est déjà plus. Dans le second cas, elle n'est rien qu'une idée abstraite, qu'on ne peut pas se représenter concrètement. Il en va un peu de cette conscience abstraite comme de la proposition « L'Homo sapiens est l'être qui pense qu'il pense ». On comprend bien que l'Homo sapiens pense au fait qu'il pense, mais on ne sait pas à quoi il pense concrètement. Penser au fait de penser ne semble pas aussi visuel que penser à la pomme appétissante qui se trouve au milieu de mon jardin.

Aux vues des périlleuses impasses qui flottent sur sa tête, l'on aurait beau jeu de discréditer la conscience. Certains scientifiques affirment d'ailleurs que la conscience est une illusion. Ce ne serait rien de vraiment subjectif en dehors de cet amas de cellules, de nerfs, de tissus — de matière morte. Chercher la présence d'une vie consciente dans le cerveau s'apparenterait à une fausse chasse au trésor. Je suis d'accord avec ce point de vue, mais uniquement si on accepte la conscience comme une réalité individuelle. L'individu se définit en effet, dans sa substantifique moelle, comme la partie indivisible d'un tout — l'individu est un citoyen au sein de la société, un être vivant dans le monde animal, un ensemble d'atomes dans l'Univers — tandis que la conscience, en tant qu'expérience subjective, est présente partout et nulle part à la fois. *Partout*, car tout ce que nous voyons, nous pensons et nous expérimentons, c'est un phénomène subjectif de conscience. *Nulle part*, car aucune chose que nous voyons, que nous pensons et expérimentons n'est la conscience objective. C'est pourquoi lorsque l'on cherche la conscience dans la matière, on ne la trouve nulle part ; et lorsqu'on cherche la conscience dans l'expérience consciente elle-même, on la trouve partout. Il s'ensuit que

la conscience ne peut pas être envisagée comme une partie du tout. Elle se tient, pour reprendre l'expression du philosophe Wittgenstein, aux frontières du tout. Elle est la limite de notre individualité.

Je ne nie pas qu'il y ait bien quelque chose de profondément individuel dans l'expérience de conscience. Cette chose individuelle, c'est le Moi qui pense, qui sent, qui éprouve. Le fait d'être un Moi autonome et digne de considération est d'ailleurs le droit le plus précieux et le plus digne d'être défendu. Mais ce Moi n'est pas la conscience dans son entier : ce n'est qu'une partie de celle-ci. Le Moi, en effet, n'existe qu'en opposition à un non-Moi. Lorsque je dis « je suis sage », « je suis en bonne santé » ou « je suis jeune », l'on rejette aussitôt les pensées qui ne sont pas sages, les sensations qui nous semblent malades et l'impression que nous avons de vieillir. Pour le dire en des termes freudiens, le Moi est une construction qui s'obtient en refoulant une partie de soi-même. Même lorsque nous nous regardons dans la glace, que nous pointons du doigt l'image de notre corps sur la surface réfléchissante, que nous nous exclamons « je suis ce corps », nous procédons à un découpage. Nous isolons en effet un objet de la conscience — la perception de ce corps — des autres perceptions existantes : celles du miroir, de la pièce et des autres objets qu'elle contient. Or, dans un certain sens, la pièce, les objets, le miroir sont aussi dans ma conscience. Ils sont dans ma conscience et existent en moi, mais ce ne sont pas le Moi à proprement parler.

Le Moi, bien sûr, ne peut pas être obtenu une fois pour toute. Il semblerait au contraire que nous passions par différentes représentations du Moi tout au long de notre développement psychologique. Ces facettes de nous-mêmes, ces persona, peuvent entrer en conflit. Elles se contredisent, se répondent, s'affirment ou se démentent ; elles s'organisent en suivant des schémas relationnels complexes. « Je est un autre », comme disait le poète. C'est pourquoi le rôle le plus important de la métaphysique, en l'état où elle se trouve aujourd'hui, consiste peut-être à distinguer le Moi du Je. Le Moi est la tentative sans cesse renouvelée de s'objectiver en tant qu'individu. Le Je quant à lui est l'évidence d'être toute chose ; que rien n'existe d'autre que de la conscience. Le Je a quelque chose d'universel. Il est en nous et en dehors de nous ; il efface les frontières entre le sujet et l'objet. Le Moi, en revanche, est individuel.

Si l'on comprend les choses en ces termes, le problème central consiste à explorer quelle relation exacte le Je (conscience subjective) et le Moi (conscience objective) entretiennent l'un avec l'autre. Cette tâche, qui sera l'objet de cet essai, consiste à déterminer les lois dynamiques permettant à la conscience de se penser, certes paradoxalement, comme une individualité. À bien des égards, ces lois s'apparentent à une lutte. Le Moi est comparable à l'homme du désert qui dispose des sacs au milieu des sables brûlants, qui érige des enceintes de tissus, construit des paravents pour se défendre de la tempête, avant de s'écrier dans le vent hurlant : « Je suis, Moi, vivant ! » Peut-être ferait-il mieux de lester ses murailles ? D'ouvrir un trou dans la

tente de toile ? De laisser passer le vent ? Peut-être alors entendrait-il ce vent lui murmurer tendrement : « Je suis vie, Je suis invisible, Je suis conscience vive. ». Car enfin la conscience n'est pas morte — oh que non ! — elle n'a fait que donner forme à ce mot de Maurice Maeterlinck :

Il est bien rare qu'un mystère disparaisse, d'ordinaire il ne fait que changer de place.