UCLouvain

Faculté de philosophie, arts et lettres
Commission doctorale du domaine CDP
Institut supérieur de philosophie

INSTITUT SUPÉRIEUR DE PHILOSOPHIE

Artificial Intelligence:
Thinking Stone or Philosopher's Stone?

Outline of a Mereological Theory of Intuitive Judgement

Membres du jury :

Peter Verdée, co-supervisor
Tom Dedeurwaerdere, co-supervisor
Bruno Leclercq, external reader
Bart Van Kerkhove, external reader
David Watson, external reader
Mark Hunyadi, president

Dissertation written by
Gwenaël G. C. Laurent

Presented for the degree of
Doctor of Philosophy

Louvain-la-Neuve
10th of November 2025

The fact that [these machines] do better than we do, does not prove that they are endowed with mind, for in this case they would have more reason than any of us, and would surpass us in all other things. It rather shows that they have no reason at all, and that it is nature which acts in them according to the disposition of their organs.

DESCARTES

Contents

Introduction **13**

Ethical motivations . . . 16
A philosophical problem . . . 19
Top-down programming . . . 21
Bottom-up programming . . . 23
Intuitive Judgement . . . 26
Artificial IQ . . . 29
Thesis Outline . . . 33
Part 1: The Problem of Judgement in Philosophy . . . 33
Part 2: Review of the Concept of Judgement . . . 37
Part 3: The Intuitive Judgement in Front of the Machine . . . 43
Part 4: Towards a General Theory of Mind . . . 45

The Problem of Judgement in Philosophy
From analytical to synthetic **53**

1 Judgement in Plato:
Origin of the Problem **53**

1.1 Socrates vs. Protagoras:
Monologues of the Soul . . . 57
1.1.1 Judgement of Experience . . . 58
1.1.2 True Judgement without justification . . . 59
1.1.3 True Judgement with Reason . . . 61
1.2 The Contemplation of Beauty . . . 63
1.2.1 Analogy of Plato's Divided Line . . . 64
1.2.2 The Pathemata . . . 65

2 Judgement in Kant :
Genesis of the Understanding **69**

2.1 The Intuitive Abstraction . . . 70
2.1.1 The Synthetic Judgement *a priori* . . . 73
2.1.2 The Limits of Platonism . . . 76

2.2 Kant vs. Plato: The Trial of Ideas . . . 79
2.2.1 The Aviary Metaphor . . . 80
2.2.2 Knowledge *a priori* . . . 84
2.2.3 The Copernican Revolution . . . 87
2.3 Sensitivity: The Man Who Feels . . . 89
2.4 Understanding: The Thinking Man . . . 91
2.4.1 Kant vs François Chollet . . . 93
2.4.2 Objective Deduction . . . 96
2.4.3 Subjective Deduction . . . 98
2.4.4 Kantian Schematism . . . 99
2.5 Reason: The Man Who Thinks He Thinks . . . 101
2.6 Beauty Without Truth . . . 103
2.6.1 The Importance of the First Critique . . . 108

3 Artificial Judgement: Can the Machine Judge? 111
3.1 Does the Machine Lack Intuition of Space? . . . 117
3.2 Does the Machine Lack Intuition of Time? . . . 120
3.2.1 The Saccade Model . . . 121
3.2.2 Objection: Auditory Chronostasis . . . 123
3.2.3 Hypothesis: Differential Perception . . . 125
3.2.4 Subjective time vs Objective time . . . 128
3.3 Does the Machine Lack Contemplative Reason? . . . 131
3.4 Does the Machine Lack Understanding? . . . 134
3.4.1 Deficiencies at the Level of the Categories . . . 136
3.4.2 General Logic vs Transcendental Logic . . . 139
3.5 Is the Machine an I? . . . 141
3.6 Does the Machine Think Differently? . . . 142

Revision of the Concept of Judgement
Toward a Mereological Theory 149

4 The Forgetting of the Understanding:
Dead of Kantian Judgement 149
4.1 The Crisis of the Foundations of Mathematics . . . 155
4.2 The Book of God and the Parallel Worlds . . . 159
4.2.1 Spherical and Elliptic Geometries . . . 163
4.2.2 Hyperbolic Geometries . . . 165
4.3 An Imaginary Geometry . . . 167
4.3.1 Lobachevsky's Constructivism . . . 169
4.3.2 Poincaré's Consensualism . . . 170
4.4 Kant Turns in His Grave . . . 171
4.4.1 Is Space a Thing in Itself? . . . 171
4.4.2 Is Space a Construction? . . . 173

4.4.3 Are Mathematics Synthetic *a priori*? . . . 174

5 Digital Rationality: The Birth of the Computer 179

5.1 The Dream of a Universal Calculator . . . 181
5.2 Theories of Provability . . . 184
5.2.1 The Barber Paradox . . . 185
5.2.2 Cantor's Paradox . . . 187
5.2.3 Circumventing the Paradoxes . . . 189
5.2.4 Gödel's Theorems . . . 191
5.2.5 Truth is a Sieve . . . 194
5.3 Theories of Computability . . . 198
5.3.1 Formalism of the Turing Machine . . . 200
5.3.2 The Halting Problem . . . 209
5.3.3 Rice's Theorem . . . 212
5.3.4 Oracles and Cybernetics . . . 214

6 Top-Down Approach: Expert Systems 219

6.1 The General Problem Solver . . . 225
6.2 Propositional Logic . . . 229
6.2.1 Alphabet . . . 230
6.2.2 Grammar . . . 231
6.2.3 Semantics . . . 232
6.2.4 Formulas . . . 234
6.2.5 Axiomes . . . 234
6.3 Expert Systems . . . 237
6.4 Conjunctive Normal Form . . . 239
6.5 Satisfiability Problem (SAT) . . . 242
6.5.1 Complexity of 2-SAT . . . 244
6.5.2 Complexité de 3-SAT . . . 246
6.5.3 Complexity of CNF-SAT . . . 248
6.5.4 Spatial Complexity . . . 248
6.6 NP-Complete Problems . . . 249
6.7 Epistemological Critique: Between Platonism and Logicism . . . 254

7 Bottom-Up Approach: Neural Networks 259

7.1 The Biological vs Formal Neuron . . . 260
7.1.1 The Formal Neuron . . . 263
7.2 Hebb's Rule: Birds of a Feather Flock Together! . . . 266
7.2.1 Isolated Neuron . . . 266
7.2.2 Single-Layer Network . . . 271

7.2.3 Multi-Layer Networks . . . 271
7.3 Deep Learning: Myth or Reality? . . . 273
7.3.1 Genetic Programming . . . 277
7.3.2 Unsupervised Learning . . . 279
7.4 Epistemological Critique:
Between Instinct and Reasoning . . . 281
7.4.1 The Achilles' Heel of Neural Networks . . . 282
7.4.2 The Neural Network vs. the Biological Brain . . . 286

8 The Blind Spot of AI:
Sketch of a Holistic Theory of Judgement **293**
8.1 Intuitive Concepts . . . 297
8.2 A Tricameral Theory of Cognition . . . 301
8.2.1 Rationality: Abstract Concepts . . . 302
8.2.2 Sensation: Statistical Percepts . . . 304
8.2.3 Perception: Intuitive Concepts . . . 306
8.3 Critique of Sequentiality:
The Intuition of the Whole . . . 310
8.4 Critique of Univocity:
The Intuition of the Infinite . . . 312

Intuitive Judgement in Front of the Machine
Extension of a Turing Machine **321**

9 The Dilemma of Intuition **321**
9.1 The Intuition of Infinity and the Intuition of the Whole . . . 324
9.2 Alan Turing's Objection . . . 328
9.3 Intuition according to Hubert Dreyfus . . . 331
9.3.1 Objections to Dreyfus's Position . . . 336
9.3.2 Challenging the status quo . . . 337
9.4 First Solution: Mysterianism . . . 338
9.4.1 Ignoramus et ignorabimus . . . 339
9.4.2 McGinn's Stance . . . 340
9.4.3 Main Objections . . . 342
9.4.4 Is Intuition a Mystery? . . . 344
9.5 Second solution : Intellectual Intuition . . . 346
9.5.1 The Contemplation of Ideas . . . 346
9.5.2 Kantian Critique . . . 349

10 Chuck Norris's Machine **353**
10.1 Functions and Models . . . 354
10.1.1 SNARC Effect: The Intuitive Structure of Arithmetic 356
10.1.2 Towards a Model of Transfinite Computation . . . 358

10.1.3 The Infinite and Undecidable Problems 361
10.1.4 The Whole and NP-Complete Problems 362
10.1.5 Responses to a Preliminary Objection 363
10.2 The Transfinite Turing Machine 365
10.2.1 Formal definition 367
10.2.2 Restrictions and Extensions of a Transfinite Machine . 373
10.2.3 Computational Power 375
10.2.4 Minimal Representation 381
10.2.5 The k-tape Machine 384
10.3 Comparison with existing littérature 386
10.3.1 Comparison with a Hierarchical Oracles Machine . . . 386
10.3.2 Comparison with a Continuous Turing Machine 388
10.3.3 Comparison with the Infinite Time Turing Machine . 395
10.3.4 Universal Machines and the Generalized Halting Problem 398

11 Generalised NP-complete problems **405**
11.1 Classical NP-Complete Problems 407
11.2 Problem 3-SAT$_{\aleph_0}$ 411
11.3 Geometric Representation of the Problem 411
11.4 Solution Method 417
11.5 Program for a Transfinite Machine 423
11.6 Strategies Concerning $P = NP$ 426

Towards a General Theory of the Human Mind What Neuroscience Tells Us **433**

12 The Mind-Body Problem: Perspectivist Mysterianism **433**
12.1 The Unsustainable Unity of Body and Mind 436
12.2 Hypothesis 1: Living Consciousness and Dead Consciousness 439
12.3 Hypothesis 2: The Paradox of Consciousness 442
12.3.1 Consciousness is Unrepresentable 443
12.3.2 Externalist and Internalist Chiasms 446
12.3.3 Links with Logical Paradoxes 448
12.4 Hypothesis 3: Dual-Aspect Monism 455
12.4.1 First Analogy: Space and Extended Bodies 458
12.4.2 Second Analogy: The Incompleteness Theorems 465
12.4.3 Third Analogy: The Principle of Complementarity . . 466
12.4.4 Fourth Analogy: Dante's Hypersphere 469

13 Critique of Encephalocentrism **473**
13.1 Is Consciousness in the Brain? . . . 475
13.2 Functional Segregation and Neural Plasticity . . . 477
13.3 Neuroscientific Spinozism . . . 479

14 Gerald Edelman's Darwinian Theory **485**
14.1 Edelman's Critique of Dualism . . . 486
14.2 Edelman's Critique of Computationalism . . . 488
14.3 Theory of Neuronal Group Selection (TNGS) . . . 491
14.4 Edelmanian Theory of Consciousness . . . 494
14.5 Metaphysical Chiasms and Epistemic Principles . . . 497

15 The Integrated Information Theory **503**
15.1 Strategy of Structural Equivalence . . . 506
15.2 Axioms of Phenomenological Experience . . . 509
15.3 Comparison with Other Axiomatic Systems (Aleksander & Dunmall) . . . 515
15.4 Physicalist Postulates . . . 518
15.5 Comparison with Shannon Information . . . 522
15.6 Ontology: Between Eleaticism and Spinozism . . . 524
15.7 Metaphysical Chiasms and Epistemic Principles . . . 527

16 Foundations of a Mereological Theory of Mind **531**
16.1 Critique of Cognitive Empiricism . . . 532
16.1.1 The Problem of the Genesis of Consciousness . . . 533
16.1.2 Dual-Aspect Investigation Method . . . 536
16.2 The Isolated Brain Experiment . . . 538
16.2.1 I think, therefore I link . . . 542
16.2.2 The Self and the Unconscious . . . 548

Conclusion **553**
Perspectivist Mysterianism . . . 555
Strong and Weak Thesis . . . 556
ChatGPT: An Alchemy of Knowledge . . . 558
Artificial Apocalypse or Last Judgement? . . . 560

Bibliography **565**

Appendix **583**

Appendix 1: A Short Story Written by ChatGPT **583**
Story Guidelines . . . 583
To Your Pens! . . . 583
A Web of Unlikely Encounters . . . 584

The Weaver of Whispers . . . 586

Appendix 2: Detailed Hypotheses of Gödel's Theorem **591**

Appendix 3: The Church-Turing Thesis **593**
Effective Method . . . 593
Turing's Theorem . . . 595

Appendix 4: Cantor's Paradox **597**
How to Define the Size of an Infinite Set? . . . 597
Approach 1: Inclusion . . . 598
Approach 2: Density . . . 599
Approach 3: Cardinality . . . 601
Counterintuitive Properties of Infinity . . . 604
Cantor's Diagonal Argument . . . 605
The Set of Subsets . . . 606
Generalisation to Any Set . . . 607
Cantor's Paradox . . . 608

Appendix 5: Transfinite Numbers **611**

Appendix 6: Shannon's Information Theory **613**
Historical and Mathematical Perspective . . . 614
Hartley and Wiener's Approaches . . . 614
Shannon's Approach . . . 615
Axiomatic System . . . 617
Shannon's Measures . . . 618
Joint and Conditional Entropies . . . 618
Relative Entropy . . . 619
Mutual Information . . . 620

Appendix 7: Essay - Is Consciousness Dead? **621**
Présentation . . . 621

Introduction

At the end of the century, one will be able to speak of machines thinking without expecting to be contradicted.

Alan Turing

Perhaps not a year goes by without a leading figure in the tech world, suddenly struck by a lightning-bolt idea, making headlines. *Eureka!* they proclaim on social media. *Consciousness is born!* they exclaim—only to be dismissed from Google shortly thereafter[1]. These Archimedes of artificial intelligence take pride in having birthed a machine endowed with thought, sensation, emotion, and creativity.

Despite the unwavering enthusiasm such declarations elicit, metaphysical claims about machines too often amount to little more than a clever intellectual sleight of hand. Nor was it necessary to await the dawn of our century—with its parade of Terminator-style chatbots—to see such assertions emerge. Perhaps you recall the Mechanical Turk, the machine built by Johann Wolfgang von Kempelen in 1770? A contraption of wires and gears, fashioned in the likeness of a turbaned Turk, it, too, made headlines. It played chess—and frequently won—against an array of famous opponents, including Napoleon Bonaparte, Catherine II of Russia, and Benjamin Franklin. Yet the crowd remained oblivious to the truth: concealed within was a living, breathing human, who manipulated the automaton's arms, hands, and fingers through a system of ingenious pulleys.

Drawing from this amusing example, one might legitimately wonder: do today's algorithms—credited with imagination, creativity, even aesthetic

[1] Blake Lemoine was dismissed on July 22, 2022, following allegations that the chatbot he was working on should be considered a full-fledged person. Indeed, he had made public a conversation he had with LaMDA (Language Model for Dialogue Applications), thereby violating confidentiality rules. Some observers, however, pointed out that Google disapproved of the moral implications embedded in Lemoine's assertions. In 2020, for instance, Margaret Mitchell and Timnit Gebru were both dismissed after warning about the ethical dangers posed by AI (Courier International, extitIntelligence artificielle. Google renvoie l'ingénieur qui disait que son IA était douée de conscience, July 23, 2022).

sense—also conceal the invisible hand of human intelligence? Behind DALL-E's and Midjourney's romantic frescoes, AIVA's baroque symphonies, and ChatGPT-4's Parnassian verses, does one not find, in the end, the statistical compilation of millions of artistic human productions?

Wherever we turn, we find ourselves caught between two opposing sentiments: on the one hand, an entirely legitimate admiration for AI's revolutionary statistical prowess; on the other, a certain disappointment in its apparent lack of real understanding. Put differently, a chasm seems to be widening between artificial intelligence's technical feats and our still-superficial grasp of human thought itself.

From a philosophical standpoint, asking whether a machine can think is a convenient way of probing what we mean by "thought," "consciousness," and "intelligence." Failing to resolve these fundamental mysteries—perhaps insoluble ones at that—our contemporaries risk taking for granted a double assumption: (1) that the brain functions like a sophisticated computer and (2) that mental life is entirely produced by the biological brain. If these assumptions, plausible at first glance but by no means self-evident, are accepted uncritically, then every advance in computer science will appear to bring us closer to a better neurobiological understanding of the human mind—so much so that the dominant methodology in artificial intelligence will eventually overshadow the distinct approaches of neurobiology and psychology.

Shall I mention that the first artificial intelligence conference, held at Dartmouth in 1956, conjectured that "all aspects of learning and other features of intelligence can, in principle, be so precisely described that a machine can be made to simulate them"? It brought together researchers such as Claude Shannon, Nathan Rochester, Ray Solomonoff, Allen Newell, Herbert Simon, Marvin Minsky, and John McCarthy. The official communiqué, a manifesto of sorts, unmistakably foreshadowed the field's contemporary evolution:

> *An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.*

This notion that the human mind operates like a computer is what I call the *reductio ad computatum* (the reduction to computation). In its current state, it remains a mere methodological hypothesis. Though it resonates with the general public, definitive arguments are difficult to establish. Throughout this dissertation, I will attempt to pinpoint the problems that this philosophical stance—if left unchallenged—risks leading us into.

To critique reduction *ad computatum*, one must first understand what contemporary algorithms are capable of, in order to highlight, in a second step, their limitations. The starting point of this work will thus be to rec-

ognize two undeniable capacities of artificial intelligence: that of analysing and that of synthesizing information. What exactly do I mean by this?

Analysis, in its essence, consists of reasoning logically. Beginning with a set of premises, it involves making deductions and advancing, step by step, toward a conclusion. Is this not, in substance, what Blaise Pascal's first calculating machine already accomplished? If we focus solely on results, there is undoubtedly something in algorithms that resembles—and often even surpasses—human reasoning.

Synthesis, on the other hand, entails summarizing information and extracting something new from it. It involves a dual movement. The first, centrifugal, simplifies, reduces, abridges. The second, centripetal, fuses, unites, hybridizes. Is this not precisely what ChatGPT does when it summarizes the contents of a text? Would it not be intellectually dishonest to refuse our neural networks the recognition of a certain form of creativity, of branching thought, and of generative capacity?

The residual question is whether current AI paradigms, beyond their abilities for analysis and synthesis, are capable of *judgement*. Judgement is, after all, an aptitude that humans generally display. It is the ability to generalize and abstract information in order to distinguish truth from falsehood, to deliberate, and to make well-founded decisions. It is the pivotal point between analysis and synthesis, between reasoning and experience.

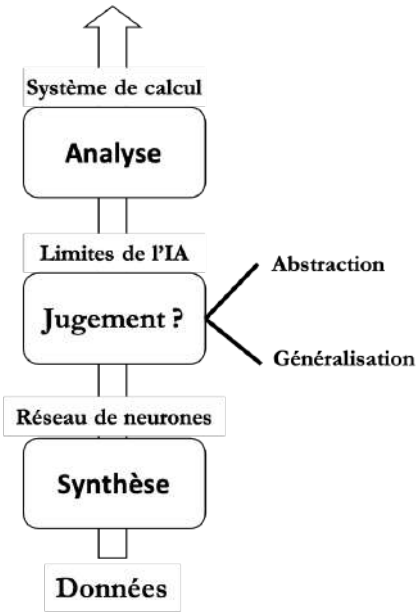


Figure 1: Judgement functions as the intermediary level between analysis and synthesis.

Within analysis, *judgement* consists in determining whether the initial

premises are true and whether the result of a calculation makes sense. It involves attributing meaning to the terms of discourse so as not to lose touch with the reality designated by the variables of the computation. This is what you do when, after solving an equation and representing each variable in the problem with alphanumeric symbols, you then ask yourself whether the values obtained make sense. You contextualize your reasoning, interpret the solution, and critique it. Is it a good approximation? Was one of my hypotheses false? Do the physical principles applied correspond to the relevant context? Judgement is thus what allows us to manipulate abstractions without losing sight of the reality they designate.

Within synthesis, *judgement* consists in reducing the diversity of information to concepts. It is thus a capacity for generalization. You employ it, for instance, when attempting to describe Frida Kahlo's style by referring to "figurative art," "the expression of suffering," or "magical realism." These expressions have meaning for you and (hopefully) for your interlocutors. However, you would be hard-pressed to explain exactly what these concepts designate using words and precise definitions. You can certainly provide clarifications on what you mean by "magical realism," but not a definition with the irrefutable clarity of a mathematical formula. There are two reasons for this. The first is that your memory is not vast enough to encompass, in a single instant, all the figurative paintings that fall under magical realism. The second is that the concept of "magical realism" may not be merely the nominalist collection of paintings bearing that label, but rather something more—a possibility? A structure? A set of properties? This is what must be examined.

The question raised by this thesis can thus be formulated as follows:

> *Does artificial intelligence possess this capacity for judgement, which appears to be the exact middle ground between the faculty of synthesis and the faculty of analysis? If not, can it understand and manipulate concepts such as "magical realism"?*

Ethical Motivations

The question of whether a machine possesses the faculty of judgement seems all the more pressing today, as an increasing number of decisions are entrusted to algorithms. When a bank must grant a loan to a client, when an educational institution must assess a teacher's abilities, when a judge must estimate the likelihood of a convict reoffending, it is now possible to rely on a program's evaluation. This program processes data concerning the individual (name, address, criminal record, profession, etc.) to determine whether they statistically resemble profiles associated with good performance. The richer the algorithm's database, the more precise its evaluation will be.

However, the data we use is not necessarily fair or accurate. In the United States, African Americans are often judged more harshly than white individuals. Faced with the same crime, they receive heavier sanctions. As a result, if an algorithm is fed real-world data, it risks perpetuating an even greater severity toward these individuals. Even if sensitive variables such as ethnicity or skin colour are not explicitly included in the database, correlations can be drawn from non-sensitive data. Postal addresses, names, and first names, for instance, are often correlated with ethnicity. A program might determine that individuals named Mehdi and residing in the suburbs are statistically more likely to commit crimes.

This form of discrimination is difficult to avoid with the algorithms we use today, for three reasons. The first is that leading tech companies are not particularly incentivized—financially or legally—to invest heavily in ethical algorithms free from bias. The second difficulty lies in the data itself. Big data is so vast that a single human lifetime is insufficient to review everything said on Instagram in just one day. The third reason, finally, is intrinsic to the most commonly used programming methods: neural networks. Within these networks, it is not easy to prevent undesirable correlations between sensitive and non-sensitive variables.

lustrate the problem of algorithmic bias, I would like to examine the responses produced by an American legal assistance algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Introduced into courtrooms in 2016, this tool is far from an impartial judge. In the image below, you will see the risk assessment of recidivism assigned by the algorithm to two individuals. The two defendants, Vernon Prater and Brisha Borden, were arrested for similar offences: the former attempted to steal tools from a store, while the latter, fearing she would be late picking up her sister-in-law from school, sought to "borrow" a bicycle and a scooter left unattended on the sidewalk—before ultimately changing her mind. Unluckily for her, a neighbor, seemingly concerned with upholding public order—yet skilfully concealed behind the curtains of their living room—reported her to the police.

In both cases, the theft was estimated at around USD 80. However, Vernon was a forty-year-old white man who had previously been arrested for two armed robberies, while Brisha was an eighteen-year-old Black woman with only juvenile infractions on her record. When calculating the risk of recidivism, the system assigned Vernon a score of 3 and Brisha a score of 8. Several years later, Vernon was arrested again for an armed attack, while Borden committed no further crimes. It is plausible to believe that most human beings—except for members of the Ku Klux Klan—would not have made this error. This obviously does not mean that the algorithm is racist, in the sense of necessarily discriminating against African Americans, but it does illustrate that statistical responses may follow a logic that does not

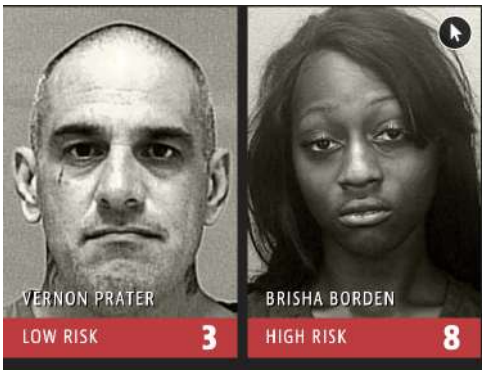


Figure 2: Algorithm assigning a numerical recidivism risk score to two defendants (2014).

coincide with the case law currently in force[2].

Algorithms have evolved considerably since 2014, yet they remain prone to errors. Do you recall that smartphone application, released in 2021, which identified a Black woman as a gorilla? Rather than being fixed, the feature allowing the recognition of primates was simply removed from the application. It is, indeed, incredibly difficult to pinpoint the root cause of such an error and correct it definitively. The neural networks used today operate in such a way that it remains, alas, difficult to provide guarantees. Even when a system performs well on its test database, it is not immune to errors when faced with real-world cases. Despite this flaw, entrusting decisions to algorithms remains convenient for two reasons.

First, computer programs benefit from an aura of objectivity. The belief is firmly entrenched that a machine—unaffected by mood swings or favouritism—operates with unwavering impartiality.

Second, delegating decisions to a machine provides a way to absolve oneself of responsibility. In the event of a dispute, it is easier to blame the machine than a flesh-and-blood human. Moreover, the law does not always clearly establish who should be held accountable for an algorithmic failure: the user? The program's developer? The company employing the machine? Those who generated or input the data?

It thus seems crucial to grasp how a machine thinks and how it decides to classify certain individuals into one category or another. If I can demonstrate that there is something in human judgement that cannot be

[2]Let us note that there is also a lack of transparency regarding the way in which the COMPAS algorithm is, or is not, used in the decisions that are rendered. This case is well documented in an article by ProPublica (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing). For a more in-depth analysis of algorithmic bias, see O'Neil's work entitled *Weapons of Math Destruction*. For a more detailed discussion of black boxes, see *The Age of Secrecy and Unfairness in Recidivism Prediction* (Cynthia Rudin, Caroline Wang, Beau Coker, 2018) or O'Neil, *Weapons of Math Destruction*

replicated by algorithms, I will hold a compelling argument against placing too much trust in machine-made decisions. But this first requires identifying the fundamental philosophical problem that artificial intelligence engineers inevitably confront when they attempt to endow machines with the capacity for judgement.

A philosophical problem

One of the fundamental problems of artificial intelligence is determining how the data available to a machine (for instance, the full set of records concerning a repeat offender) can be categorized in such a way as to form manipulable concepts—guilty, not guilty, suspicious. This ability to gather the diversity of experiential data under a single category, one that can be processed by the laws of thought, is what philosophy calls *judgement*.

The issue I have just described is not merely a technical question but hinges on an epistemological enigma: the question of the relationship between sensible things and the concepts we have of them. It is one thing to see a criminal; it is quite another to say that someone *is* a "criminal." The concept of "criminal" contains within it all possible and imaginable crimes, both those we witness first-hand and those that lurk, insidiously, in the depths of what we take to be a criminal nature. By uttering so banal a phrase as *He is a criminal!*, I set apart a particular human being from the rest of the crowd, endowing him with unity, with an individual form. At the same time, I assimilate him, by resemblance, to all other criminals throughout human history. Within my perception of this man, I isolate the presence of a tormented face, a shifty gaze, a troubled past, a lack of empathy, a weakness of character, a hidden perversity, a spirit of malice. At that moment, a relation of association and differentiation is established between this singular instant of perception (I experience something in the here and now) and a universal knowledge (I say of this experienced thing that it is a face, an expression, a voice—that of a criminal).

The issue I am raising is known in philosophy as the *problem of universals*. It asks why the mind is capable of thinking the relationship between a universal concept and a singular object, and by what laws this relationship can take place. From this fundamental problem stem many questions: Do the concepts we handle possess an independent reality, or are they merely convenient expressions, linguistic short cuts for designating the objects of the world? Would we be able to perceive an object such as a tree if we did not have the concept of a tree? Would we possess the concept of a tree if we had never been exposed to the perception of particular trees?

In computational terms, these questions about universals amount to asking whether something like *raw* data exists—data that can be collected without being subjected to any predefined structure—or whether, on the con-

trary, it is indispensable to have pre-existing categories to represent data. If the latter is true, then *raw* data does not exist; all data is *formatted*.

Two traditional answers have been proposed to address the problem of universals:

- **Empirical stance:** If one holds that objects come first and that concepts are derived only through direct experience of the world, one aligns with the *empirical* tradition. The human mind, in this view, is nothing but the result of friction, of a confrontation between the real world and our representation of it. If we have the concept of tree, wolf, house, and criminal—but not a readily available concept for distinguishing different set of subtle odors—it is because the concepts of tree, wolf, house and criminal proved to be the most adaptive for the development of the human species under the conditions of the Paleolithic era.

- **Idealist stance:** If, on the contrary, one holds that concepts are first and enable us to apprehend the world—that without them, we would stumble blindly in the dark; that the sensible world, far from being populated with distinct individual objects, would be nothing more than a patchwork, a cacophony reminiscent of the primordial cosmic soup of the universe's earliest ages—then one aligns with the *idealist* tradition. This tradition asserts that there exist *a priori* structures of thought, pure forms that render our subjective experience intelligible. Concepts, in other words, are the indispensable lenses through which we contemplate the chiaroscuro of the world.

In this thesis, I hypothesize that the tension between empiricism and idealism is present within artificial intelligence strategies. Far from being confined to the idle pursuits—already quite numerous—of philosophy, the question of how concepts are formed has given rise to two programming paradigms: *top-down* programming and *bottom-up* programming. The former begins with concepts as though they were preexisting realities and attempts to say something about the real world. The latter, by contrast, begins with objects and adapts concepts as experience unfolds[3].

[3]My use of these terms differs somewhat from how they are sometimes found in the literature. *The Perl Design Patterns Book*, an online book on the Perl programming language, employs this terminology to describe the structuring of a program's architecture through rules, distinguishing between cases where these rules apply to particular situations and those where they apply to general ones. I have, however, chosen to redefine these terms slightly so that all prescriptive rule-based approaches are classified as *top-down*, whereas the term *bottom-up* applies to Deep Learning methods, particularly those based

I will detail these two paradigms in the next two chapters. But before doing so, I would like to clarify, in a synthetic formulation, the hypothesis that underpins my approach:

Hypothèse 1. *The question of how concepts are formed is the Gordian knot of artificial intelligence. The empiricist perspective holds that concepts adapt to the world around us and evolve through experience. The idealist perspective asserts that it is the objects of the world that conform to our concepts. These two perspectives have given rise to the top-down and bottom-up programming paradigms, which will be discussed in the following chapters.*

Top-down programming

When adopting the idealist perspective, one assumes that concepts have an objective existence. They govern, filter, and orient our representation of the world. Consequently, if one seeks to program an algorithm designed to perform a certain task, the first step will be to define the concepts related to that task, only then confronting the machine with particular cases. This programming approach operates under the implicit assumption that the logical analysis of concepts is sufficient to determine truth and falsehood in real-world contexts.

I would like to illustrate this paradigm with an example drawn from science fiction. Suppose we wish to shape a moral robot akin to those that Isaac Asimov's pen has accustomed us to. We would not seek to construct the concepts of *good*, *fairness*, or *justice* by drawing from social practice or by simply imitating average human behaviors. On the contrary, we would proceed from the principle that morality is an objective reality, endowed with laws to which society must conform. Once these laws have been identified, it would suffice to encode them into a programming language. Each term of the moral law must be expressible as a sequence of binary numbers. Each decision, for its part, must be presented as the result of a calculation performed on these numbers.

The algorithm, equipped with this new grammar, can manipulate moral rules to deduce, *a priori*, which responses are morally admissible in the face of a given dilemma. For the algorithm, it is as if goodness could be reduced to mathematical proofs. Generosity, compassion, and justice become so many *theorems of morality*, susceptible to demonstration. This type of algorithm, which analyses, deduces, and reasons according to the strict laws of logic, is commonly called an Expert System. The first traces of such Expert Systems can be found in the 1950s, in domains ranging from chess to medicine, not to mention financial consulting. It is also these systems that inspired Isaac Asimov's moral laws of robotics, later adapted in the film *I,*

on statistics and self-learning. By my definition, the vast majority of programs from the 1980s would be considered *top-down*, while neural networks, for instance, belong to the *bottom-up* paradigm.

Robot, starring Will Smith, and formulated as follows:

1. A robot may not harm a human being, nor, through inaction, allow a human being to come to harm;
2. A robot must obey the orders given by human beings, except where such orders would conflict with the First Law;
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws are conceived in a wholly abstract and universal manner. The programmer did not need to refer to any particular context to establish them. It is left to the robot to determine, within each specific context, how best to apply them. This requires interpretation, deliberation, and, ultimately, judgement. Yet, these are precisely the capacities we wonder whether they can be encoded into laws. Is logical analysis sufficient to grasp a social context and understand its moral implications? Or would a robot programmed in this way be a kind of autistic caricature, a ludicrous puppet, a rigid logician incapable of comprehending non-verbal language?

The programming paradigm that begins with analytical laws and assumes that the context must conform to them is what I call the *top-down* paradigm. The advantage of top-down programming is that the algorithm functions transparently. Since the governing laws of the model are fully known, one can trace back, step by step, to the principle that gave rise to the machine's behavior. If a robot refuses to obey an order, one will know that, according to its analysis, the order endangers a human life. This is why such algorithms are generally considered *traceable* and *justifiable*.

This traceability does not mean that everything the machine does is absolutely predictable. Even though the laws are known in advance, it is not always possible to determine, in advance, what result they will yield. This kind of logical unpredictability is comparable to what is known in mathematics as chaotic behavior. Two identical sets of equations, confronted with similar contexts, may produce radically different outcomes. Anyone who has ever attempted to predict the weather or stock market fluctuations has likely experienced this first-hand. The slightest factor can significantly alter the behavior of the whole.

Asimov seems aware of this phenomenon. In his novella *Robots and Empire*, the android R. Daneel Olivaw deduces, based on the initial laws of robotics, the existence of a Zeroth Law, stating that a robot must never harm humanity. He decides to override the previous laws in favor of this higher principle. In doing so, he does nothing other than apply the ethical code he was taught with the utmost rigor. His behavior is unexpected, even

chaotic, yet it remains entirely traceable and justifiable—or at the very least, justified.

Similar surprises can arise in vastly different applications of the Moral Robot. Using the top-down method, for instance, it is possible to derive mathematical proofs that, while entirely logical and predictable, would not necessarily have occurred to a human being. A program called the *Logic Theorist*, developed by Herbert Simon and Allen Newell, was capable of proving some theorems in the *Principia Mathematica*—a work of mathematical logic renowned for its technical complexity. The program even produced proofs that were sometimes more elegant than those previously known[4]. Instead of basing its reasoning on Asimov's laws, this system relied on the axioms of arithmetic. It began with the theorem to be demonstrated and applied transformations until it reduced the problem to an accepted axiom.

However, this gain in transparency and understanding comes at the cost of a lack of adaptability. A top-down program, whether designed to resolve moral dilemmas or mathematical problems, is incapable of forming new concepts. Nor can it think beyond the contexts covered by its analytical laws. What would Asimov's robot do if the life of an animal were in danger? Or if the ethical issue at stake concerned environmental preservation? The top-down machine is a horse with blinders. It follows the tracks laid out for it, never straying from its designated path.

This is why it may be wise to ask, contrary to the idealist approach, what happens to programming when one adopts the perspective that concepts have no objectivity—that they are ideas formed through contact with experience?

Bottom-up programming

We have seen that Isaac Asimov's robots are programmed to follow a predefined moral law. They possess universal and rational concepts, which they attempt to apply to the real world. This approach aligns with the idealist perspective. However, one might rightfully wonder what would become of these robots if we were to adopt an empirical perspective instead. In such a scenario, moral good would have no objective existence. Lacking a norm inscribed in the marble of eternity, the only way to determine what is just or unjust would be through the observation of human behavior.

What kinds of strategies might such an empirical vision lead to? Had Asimov been born today, he would likely have mocked the laws of robotics, deeming them too narrow, too inflexible. Rather than programming analyt-

[4]This result was actually achieved by the Logic Theorist, the predecessor of the General Problem Solver. This software produced a proof of theorem 2.85 from the *Principia Mathematica*, which was judged to be shorter and simpler than that of Whitehead and Russell.

ical rules, he would have been content to provide the machine with as much data as possible to synthesize. He would have wagered that, after adequate training, the robot could independently identify which behaviors should be deemed "just" and which should be regarded as "unjust." To boost artificial learning, he would likely have resorted to the classic method of the carrot and the stick. Correct responses are reinforced, incorrect ones are eliminated. With each new generation of robots, one would expect to see moral performance grow as surely as a rosebush flourishing under the favor of the spring sun.

This approach is widely used in the design of chatbots. These systems can express themselves on moral issues—or any other subject to which they have been previously exposed. The first somewhat convincing conversational agent dates back to the 1970s. This program, called Eliza, processed sentences by analysing their grammatical structure: subject, verb, object, and so forth. It was designed to associate each input sentence with a statistically coherent response. As one might expect, these associations lacked the subtlety of ChatGPT-4. Nevertheless, the program managed—ironically—to pass as a caricature of a Rogerian therapist, known for rephrasing a patient's words in a pseudo-intelligent manner. Below is an excerpt from a conversation I had with Eliza:[5] :

ELIZA: How are you? Please tell me your problem.
YOU: I'm trying to describe you for a book I'm writing.
ELIZA: Do you like writing?
YOU: Yes, but not necessarily about you.
ELIZA: That is quite interesting.
YOU: What can I write about you?
ELIZA: Perhaps, in your fantasies, we are writing about each other.

Naturally, conversations with Eliza never passed the Turing test. It takes only a few minutes for an average interlocutor to realize that they are not speaking to a flesh-and-blood human but to a machine of copper and silicon. However, the statistical methods behind 1960s chatbots would benefit from several significant innovations, leading, in 2022, to the emergence of the ChatGPT we use today. These innovations include neural networks (introduced as early as the 1950s and perfected in the 1980s) and transformers (in 2017).

Neural networks were honoured by the 2024 Nobel Prize in Physics, awarded to John Hopfield and Geoffrey Hinton for their pioneering work. Their approach processes information through sophisticated statistical regressions. This regression aims to simplify received data by identifying

[5] I constructed this example by visiting a webpage that reproduces Eliza's code: `https://www.masswerk.at/elizabot/eliza.html`

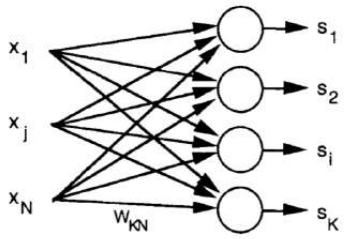


Figure 3: The neural network connects the initial neurons on the left (the problem's data) to the terminal neurons on the right (the admissible responses).

trends. Suppose, for example, you are given a scatter plot representing the number of crimes committed by individuals as a function of their income. You might try to draw a line indicating that below a certain income threshold, at least two crimes are committed on average over a lifetime. This method, though straightforward, is nothing more than a statistical approximation. It remains far removed from the fiction imagined in the film *Minority Report*, where beings with special abilities (called precogs) can predict crimes with absolute precision.

Needless to say, real-life situations are not reducible to a single axis of analysis, such as income, but must account for numerous other factors related to the defendant's criminal record: the number of prior offences, their nature, length of incarceration, and so forth. Neural networks use the same principle. Each piece of data in a problem is represented by an initial neuron: there is one neuron for a person's income, another for their age, another for the number of crimes they have already committed, etc. These initial neurons are interconnected, as well as linked to terminal neurons that correspond to possible outcomes: "the defendant will commit a crime" or "the defendant will not commit a crime." There may also be a terminal neuron for each subcategory of response: "likely criminal," "unlikely criminal," "highly unlikely criminal," and so on. The system is trained to determine which terminal neuron should activate based on the information carried by the initial neurons. This training is purely statistical. The network establishes connexions that best reflect the observed outcomes in known contexts, hoping to provide reliable responses in new situations.

The primary technical challenge posed by neural networks is adequately linking information. As long as the words in a sentence—or the data related to a criminal case—are considered in isolation, it is difficult to grasp the meaning of the sentence or determine the defendant's culpability. To refine the machine's judgement and approximate what a human does when reading a document or analysing an accused person's file, it is crucial to establish connexions between the elements of the problem and to model how they

influence one another.

It is precisely to address this challenge—the articulation of neural information—that the second major technology, transformers, was developed. In a 2017 paper by Ashish Vaswani titled "*Attention is All You Need*," a neural network model designed to analyze language was introduced. Its uniqueness lies in processing sentences not sequentially but in parallel. Instead of examining words letter by letter and sentences word by word, the algorithm scans the sentence, drawing inspiration from human attentional processes.

The key idea behind transformers is to determine how the presence of one word within a sentence influences the interpretation of adjacent words. Naturally, like most neural networks, the algorithm only works if fed a vast dataset—Wikipedia pages, websites indexed by Google, or user messages harvested through smartphone applications. The program leverages all these examples of known sentences to refine how it directs its attention within a sentence. In summary, one might say that a transformer is a predictor designed to determine, within a fragment of text, which word should come next. In this sense, it resembles traditional grammar exercises:

Pablo is ... in the classroom.

Marc did not ... the order.

The transformer's approach, based on statistical association and self-learning, falls under what I call the *bottom-up* programming paradigm. The advantage of bottom-up programming is the vast diversity of observed responses. The machine is free to discover new relationships. Its major drawback, however, is a lack of control and transparency. If you feed a chatbot conversations from social networks, there is a risk that the algorithm will become a mouthpiece for misinformation. This was, in fact, observed with the earliest versions of ChatGPT, which exhibited discriminatory biases. Some may also recall the BlenderBot 3 scandal. This chatbot, launched on August 5, 2022, by Meta, stubbornly claimed that Donald Trump was still the president of the United States—and that there was ample evidence that the elections had been rigged. Factually false as these statements were, they had something eerily prophetic about them.

Intuitive Judgement

It would seem that the top-down and bottom-up paradigms stand in direct opposition to one another. They are like the sky and the earth, like the moon and the sun, like rain and clearing skies, showers and storms. The *top-down* approach, indeed, starts from the known laws of theory, pre-encoded in advance, and seeks to derive useful conclusions from them. However, the

validity of its responses depends on nothing other than the validity of its initial premises. When one is asked, "What is the colour of Napoleon's white horse?" the answer is already inscribed in the question. Likewise, when a machine must evaluate a chess position to determine whether it is possible to "checkmate in five moves," the answer is inscribed, from all eternity, in the laws of logic. It is said that the machine produces *a priori* analytical judgements. *Analytical*, because they follow from the logical analysis of the problem's terms. *A priori*, because there is no need to refer to experience of the world.

The *bottom-up* approach, on the other hand, begins with sensory experience. It processes large sets of data to establish correlations. It thus seeks to draw connexions that are not, *stricto sensu*, contained within the definition of the terms. If it observes that generals are generally perched on white horses, it may infer that Napoleon was as well. If it observes that certain chessboard configurations generally lead to the defeat of the white pieces, it will conclude that the position is lost for White, without calculating how many moves are needed to reach checkmate. This statistical reasoning is not a logical truth. Consequently, the judgements it produces are not *a priori* analytical but rather *a posteriori* synthetic. *Synthetic*, because they lead to a truth that exceeds the definition of the problem's terms. *A posteriori*, because such a judgement requires reference to concrete experience.

The distinction I am drawing between analytical and synthetic judgements is analogous to the difference between induction and deduction. In philosophy, induction refers to the act of generalizing a recurring observation to establish a general law. It is thus a kind of *synthesis* of experience. Induction occurs when, after gazing at the Scottish meadows from a train window, you declare to your fellow passenger, "In this country, all sheep are white." You may very well be justified in summarizing your observation of the pastoral landscape in this way, but it would be more prudent to say, "All the sheep I have seen through the window are white." For if the silhouette of a black sheep were to be spotted by your neighbor, looking out from the opposite window, you would be forced to revise your judgement and say, not without a hint of British irony: "In this country, there exists at least one sheep of which one side is not white."

In contrast to induction, deduction is reasoning based not on the observation of recurring patterns but on the laws of logic. It thus provides certain truth, akin in this respect to *analytical* reasoning. To offer a simple example, deduction is what Inspector Rouletabille employs when, upon noting that a crime has been committed in the Yellow Room and assuming that there is no exit from this room, he concludes that the culprit cannot be among those who were outside at the time of the events. The justification for this reasoning lies in the fact that no one can be simultaneously *inside* the room and *outside* it. To admit that the culprit possesses the gift of ubiquity would not only contradict worldly experience; it would also violate the well-known

principle of contradiction, which states that one cannot simultaneously have a thing and its opposite.

Both deduction and induction are within the reach of algorithmic processes. Indeed, as long as we confine ourselves to logical reasoning, human beings are no match for the more or less sophisticated variations of that calculating machine known as the Expert System. Likewise, as long as we limit ourselves to establishing statistical correlations, it is difficult to compete with the high associativity of neural networks. The question that arises, then, is whether there exists another faculty besides induction and deduction—one that stands, so to speak, between purely *analytical* and purely *synthetic* judgements.

In this book, I wish to discuss two human faculties that still resist our algorithms: what I call abstraction and generalization. Concerning *abstraction*, first of all, one may ask how a machine can establish a link between rational concepts, logically defined, and statistical categories obtained through data analysis. If a neural network sees many squares, it is certainly capable of recognizing new squares, but can it, for that reason, provide the logical definition of a square? Can it state that it is a quadrilateral with four sides? And how, then, would it define a quadrilateral? Could it rediscover Euclid's axioms solely through the statistical analysis of geometric figures? This aptitude consists in abstracting the diversity of experiential data in order to reduce concepts to their minimal units.

This first type of problem leads to a second difficulty: that of *generalization*. Bottom-up algorithms often rely on the massive use of big data. Every day, millions of pages of text are fed to ChatGPT to improve its performance. The resulting algorithm behaves like a pantagruelian being that devours, fragments, and incorporates the knowledge of humanity. Yet it is legitimate to ask what would become of machine intelligence when confronted with a limited number of data points. Is the human mind not capable of reasoning with an ultimately very small amount of data, compared to the machines used today?

This concern has led white-collar technocrats to argue that it is high time to focus not only on big data but on quality data. I would like to illustrate this with an example. If I show you three shapes you have never seen before and then ask you to choose from among four new shapes the one that most resembles the first three, it is likely that you will have no difficulty responding. This ability is a key aspect of what is known as fluid intelligence. Unlike so-called crystallized intelligence, which reflects general knowledge and educational level, fluid intelligence seeks to reason without significant prior knowledge. It wipes the slate clean and constructs concepts step by step.

The problems of abstraction and generalization are intrinsically linked. Abstraction is the ability to strip experience of its disparities to reduce it to simple logical forms. Generalization, on the other hand, consists in reasoning

from a small sample of data. In both cases, the mind seems to manipulate intuitive concepts. These concepts are peculiar in that they always overflow the logical definition given to them.

A simple example of an intuitive concept is that of a "tree." Any normally constituted human being can identify a tree in the middle of a meadow, yet they would struggle to provide an exact definition of a tree. A tree is not merely "a trunk with roots and branches." Many plants meet these criteria without being trees (a tomato plant, bamboo...). Moreover, there are trees with multiple trunks, just as there are trees without branches. The intuitive concept of "tree" does not derive its intelligibility from its linguistic definition but seems to stem from a capacity for practical judgement.

Put differently, the human mind structures information by reducing it to simple objects: chairs, tables, squares, etc. The machine, on the contrary, structures information by breaking it down into a set of disjointed points: pixels, time instants, etc. It then attempts, after the fact, to reconstruct concepts by linking these points together. It is not certain that this bottom-up approach can grasp the complex, intuitive relations that govern the understanding of the whole. In the following pages, I will show that there exist abstraction and generalization tasks, devised by François Chollet, on which our best algorithms still struggle to provide an answer—tasks that are nevertheless effortless for humans[6]. The aim of my approach, then, will be to put the following thesis to the test:

Thèse 1. *Neither top-down nor bottom-up programming is capable of forming intuitive concepts in their own right, but only of reasoning about definitions or performing statistical associations. There exists a type of judgement that is reducible neither to pure analytical judgements nor to pure synthetic judgements, but that proceeds through abstraction and generalization.*

Artificial IQ

This difficulty that I have pinpointed—namely, the struggle of algorithms to abstract and generalize their reasoning—is well known in the field of computer science. It has even been the subject of highly specific test batteries. One notable example is the ARC protocol, created in 2019 by François Chollet, which consists of a series of generalization problems akin to Raven's matrices (see Figure 4). In this protocol, the machine is presented with two pixel grids: an input grid A and an output grid A'. It must determine what transformation was applied to image A to produce image A'. The same operation is then performed with grids B and B', followed by C and C'. Based on these three examples alone, the machine is then asked to transform an image D into D'.

[6] Chollet reports that, on average, nearly 85% of problems are solved by a human being (see *ARC-AGI* on `https://arcprize.org/arc`.

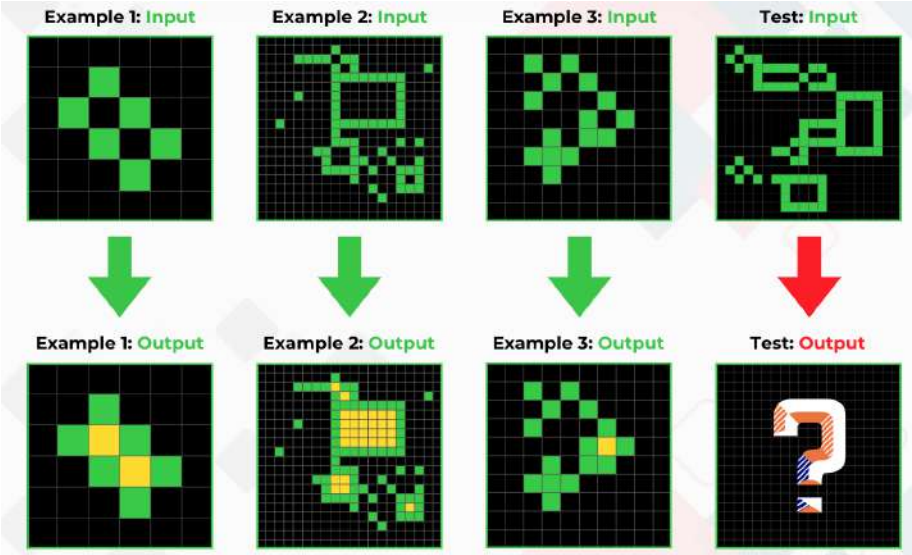


Figure 4: Here is an example of an ARC test. Given an input, the machine must provide the correct output. It receives three examples to understand what it must do (source: https://lab42.global/arc/).

The key difference between the problem I have just outlined and those typically tackled by artificial intelligence is that very little data is available for solving the Raven matrix. The machine must perform based on only three examples. This stands in stark contrast to big data-driven software like ChatGPT, which is trained on more than 800 GB of data! In traditional benchmarking, the sheer volume of data is so exorbitant—so nearly Promethean—that the entire challenge lies in sorting, classifying, and discarding data as needed. With artificial IQ tests, on the other hand, the focus is on *quality data*. What matters is not the quantity of data but rather the quality of the examples presented and, above all, the algorithm's capacity to generalize its reasoning.

The images to be analysed most often consist of simple geometric shapes, made up of a small number of pixels, subjected to transformations such as rotations, colour changes, zooms, pixel additions, and so forth. According to François Chollet, these exercises engage a number of key cognitive functions, such as those listed below[7]:

- **Object persistence:** The ability to recognize and manipulate objects without losing their identification due to a change in orientation

[7]These properties are outlined by the author on the website dedicated to ARC tests (https://arcprize.org/arc). For further details, the reader is invited to refer to his publication: François Chollet, *On the Measure of Intelligence*, November 5, 2019.

or context.

- **Goal-directed behavior:** The ability to attribute intentions or goals to certain objects.

- **Arithmetic:** The ability to sort objects based on shape, colour, appearance, or movement, relying on addition, subtraction, and comparison.

- **Geometry and topology:** The ability to perform spatial manipulations on objects, such as rotations, orthogonal symmetries, translations, deformations, repetitions, associations, etc. Notably, it is also possible to determine the distance between two objects.

It is evident that these cognitive operations closely resemble what I had in mind when discussing intuitive judgement. Essentially, they involve categorizing sensory information in such a way that it becomes manipulable—associating a rational geometric form with disparate data. A study has even shown that, on average, humans successfully complete 87

It is therefore reasonable to think that the problem of generalization poses fundamental challenges to the development of artificial intelligence. The observed performance gaps between humans and machines may in fact reveal a deficiency in the exercise of understanding itself—a fundamental limitation in the very act of applying a rational concept to sensory reality. This is the hypothesis I wish to explore in this thesis.

However, one must not entirely rule out the possibility that the mixed results of current artificial intelligence methods on ARC exercises stem simply from the novelty of these problems. Until now, programmers have relied heavily on *benchmarks* (standardized datasets) to train their algorithms for specific tasks, such as image analysis, conversational robots, or search engine development. This methodology has proven so effective that designers have had little incentive to create algorithms capable of reasoning in situations with limited data.

It is therefore not unthinkable that algorithms will improve rapidly on ARC tests in the coming years[8]. The history of artificial intelligence has

[8]However, I would like to point out that people tend to underestimate the time required to overcome a difficult task. In 1958, Herbert Simon predicted that it would take ten years for a chess program to defeat the world champion, yet this did not happen until 1997. This phenomenon is encapsulated in Hofstadter's Law, which essentially states: *"It always takes longer than you expect, even when you take into account Hofstadter's Law."* See HOFSTADTER in *Gödel, Escher, Bach: An Eternal Golden Braid*.

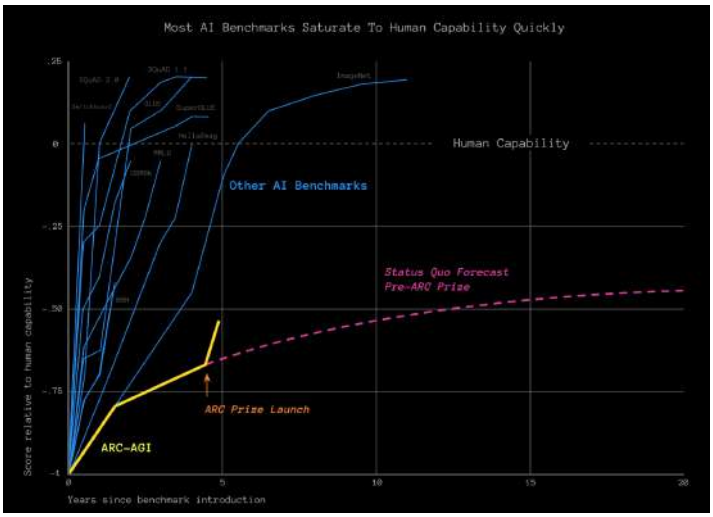


Figure 5: Most algorithms eventually surpass human abilities in various tasks. Progress on ARC tests followed a rather flat asymptote—until a USD 1.1 million prize was allocated to the challenge (source: https://lab42.global/arc/).

repeatedly demonstrated the peril of confidently declaring, *"Here is what machines still cannot do!"* Need we be reminded that AlphaGo managed to defeat Lee Sedol in 2016 without any prior database? Why, then, should Raven matrices à la François Chollet not meet the same fate?

I do not deny this possibility. My point, rather, is that a truly intelligent algorithm—one that not only passes François Chollet's test in practice but also in spirit—should not resemble a cold, soulless specialist, a technician of Raven matrices, dazzling when faced with a familiar problem yet utterly incapable of applying its formidable methods outside the narrow grooves in which it was trained.

What we fundamentally aspire to—the dream, the nightmare, the Pygmalion fantasy—is the creation of a machine capable of discerning abstract forms across diverse and varied contexts. Moreover, François Chollet specifies that a general artificial intelligence should also be able to transfer its conclusions and methods to a new environment[9]:

> *AGI (Artificial General Intelligence) is a system capable of efficiently acquiring new skills outside of its training data. This means that it is able to adapt to an environment it has never*

[9] Translated by me from English: *"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."* See the official ARC test webpage: `https://arcprize.org/arc`.

> *encountered before and that its creators (developers) did not anticipate.*

Thesis Outline

The thesis is divided into four parts. Though they follow a logical and chronological thread, these parts are designed to be self-contained. You are therefore free to explore the sections that interest you most, to skip passages, to return at your leisure—to navigate these chapters as though wandering the corridors of a clever labyrinth. The four parts are as follows:

1. The Problem of Judgement in Philosophy:
   From the Analytic to the Synthetic
2. Revisiting the Concept of Judgement:
   Toward a Mereological Theory
3. Intuitive Judgement Confronting the Machine:
   Extending the Turing Machine
4. Toward a Dual-Aspect Monism:
   What Neuroscience Has to Say

I will summarize here the purpose of each of these parts. But first, I would like to draw your attention to an important point: the first chapter of each part always provides a summary of what has been said previously, as well as what is yet to come. These chapters thus serve as relay points. There, you will find a moment of respite, a touch of humor, and a few prophecies about the future course of the book. If your reading experience resembles more that of a detective novel than the soporific recitation of a school manual or an encyclopaedia, I will consider myself to have at least escaped the dreadful accusation leveled by Johann Wolfgang von Goethe against the intellectual productions of his time:

> *Some books seem to have been written not to instruct us, but merely to show that the author knew something.*

Part 1: The Problem of Judgement in Philosophy

In the first part of this thesis, I will begin by analysing how the concept of *judgement* is defined within the philosophical tradition. I will suggest that, as early as Plato's time, one can already observe a divide between analytic and synthetic judgements. Greek philosophy, after all, was born from

the desire to elevate knowledge—entangled in sensory observations—toward greater rationality. It posits, on the one hand, the existence of truth founded on experience, and on the other, truth grounded in ideas. Socratic dialectic, in this regard, can be seen as an attempt to elucidate—through the sheer power of language—the essence of courage, science, love, wisdom, art, and the many other concepts we use in daily life.

"If we were asked about something trivial and commonplace—say, what mud is", Socrates explains in a dialogue opposing him to Theaetetus, "and we answered that there is the mud of potters, the mud of oven builders, the mud of brickmakers, would we not be ridiculous?" With this statement, the Athenian philosopher clearly expresses the idea that something as simple as mud is not merely the statistical sum, the crude synthesis, of all the possible and imaginable forms of mud. Mud is "earth mixed with water". It is an idea endowed with a rational definition, an immutable essence. It must not be confused with its various, necessarily contingent, appearances as they manifest in sensory experience.

What is interesting is that Plato does not settle for a strictly binary division, opposing the sensible world (synthesis) to the intelligible world[10] (analysis). He also envisions the existence of a form of judgement situated precisely between the heavens of rational definitions and the solid ground of experience. He calls this the "true judgement without justification". What he means by this is a certain capacity to form the concept of what is "hard", "soft", "cold", "hot" or "muddy", by gathering under a single linguistic unity all the diverse things that are hard, soft, or cold—yet without providing a rational definition of them. In this thesis, I will attempt to show that this judgement without justification is the embryonic form of intuitive judgement, which lies at the heart of my philosophical inquiry.

Although Plato devotes several passages to this "judgement without justification", he does not attribute it to any particular human faculty. Let us clarify this point. Unlike rationally justified judgement, which belongs to reason, and unlike experience, which belongs to sensation, intuitive judgement is nothing more than a general aptitude of the human soul. This is what Plato conveys through the voices of his characters: "I do not know what to say, except that, in my opinion, there is no special organ for these [judgements], as there is for others: it is the soul itself, and by itself, that, in my view, examines the common notions in all things.[11]"

It is as if Plato acknowledged the existence of a form of judgement located between experience and reason but failed to grant it the importance it truly deserved. Could this striking omission stem from his ideal of rationality? Socrates' obsession, in every dialogue where he appears, remains

[10] It should be noted that the expression "intelligible world" does not appear as such in Plato's writings but is of more recent origin. However, the expression "intelligible place" (*topos noetos*) can indeed be found.

[11] *Theaetetus*, 186b-186c.

unchanged. Whether discussing virtue, the art of leadership, or the proper way to siphon multiple amphorae of wine, his discourse relentlessly seeks to reduce all the intuitive concepts we form about the world to rationally argued concepts. Not only is mud "earth mixed with water", but science itself is "true judgement accompanied by reason".

There is a movement, in the early stirrings of Western rationality, that seems to have bracketed the legitimacy of intuitive reasoning, understood as the faculty of judging truth without relying on either formal logic or mere statistical laws. This movement, extended by modern science, is what I will call in this thesis "The Forgetting of Understanding". The term "understanding" in this expression refers to the act of grasping something through intelligence, of being able to judge it, without necessarily possessing a purely rational explanation for it. It is an intermediary between simple sensory perception (hearing something) and logical-deductive reasoning (comprehending something)[12].

At the end of this first part, I will show how Immanuel Kant's *synthetic a priori judgement* inherits this historical tension between analytic judgement (from reason) and synthetic judgement (from experience). Furthermore, I will use Kantian theory as a starting point to question the limitations of artificial intelligence. The aim will be to identify a first essential fracture between human cognition and the functioning of our current algorithms: the way information is structured.

I will suggest that algorithms break information down into irreducible units (pixels, bits...) and then attempt to associate these minimal units according to statistical laws in order to identify structures (recognizing the image of a tree, understanding the meaning of a sentence, etc.). Humans, by contrast, analyze information as a whole; rather than breaking it down into indivisible pixels, they perceive it through pre-existing structures. These structures are *a priori* in the precise sense that, in their absence, perception itself would be impossible. In other words, it is because we already possess the concept of square, of triangle, of space, of table, that we can represent to ourselves a square table in the middle of a room.

I will illustrate this phenomenon with several perceptual illusions, including Kanizsa's illusion and Ponzo's illusion. In Kanizsa's illusion (Figure 3.2), sectioned spheres are arranged in such a way as to suggest the existence of a triangle. No segment of this triangle is explicitly drawn, yet the human

[12]It is worth noting that the faculty of *Understanding*, in French philosophical texts, is translated as the faculty of *Entendement*. Interestingly, in Old French, *entendement* meant *understanding*, but this sense has fallen into disuse. Today, the French verb *entendre* means that one can *hear* a noise or a sound, but it no longer signifies the understanding of mathematics or poetry. This linguistic omission of the word *entendement* as a form of *understanding* mirrors the philosophical oblivion of intuitive judgement itself. It contributes to the modern scientific representation of the world, which holds that all truth can be reduced solely to rational calculations.

mind cannot help but perceive—somewhat disturbingly—that the triangle is indeed there. This illusion illustrates the idea that information is not processed in a purely objective manner, as an algorithm would, by associating each point in space with a coloured pixel; on the contrary, the image as a whole appears to us in accordance with certain pre-existing structures. It is because we already possess—prior to contemplating the image—the concept of a triangle that we see a triangle. One might therefore argue that the segmentation of the image depends on our *a priori* concepts.

Figure 6: In Kanizsa's illusions, a geometric figure emerges from the white of the page. From a Kantian perspective, this illustrates that it is the sensible world which conforms to our categories, rather than the reverse.

In Ponzo's illusion (Figure 3.4), something slightly different occurs. Two horizontal lines are depicted. Though they are strictly identical in length, the upper one appears longer. This optical illusion is created by the presence, in the background, of two additional lines—suggesting the existence of a perspective. If you remove these background lines, the illusion vanishes. Unlike an algorithm, the human brain cannot avoid being deceived by the illusion. This suggests that our perception does not merely examine objects individually but is influenced by the spatial arrangement of objects relative to one another. A global phenomenon in perception seems to resist the programming logics that currently govern image analysis. Moreover, this global perceptual phenomenon is shaped by our pre-existing concepts of space and three-dimensionality. Notably, the illusion is weaker in children than in adults and is also diminished in individuals with schizophrenia[13].

[13]For the effect of age, see 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. For schizophrenia's impact, see 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. Note that the studies I mention here do not specifically concern the Ponzo illusion but rather the Chubb illusion, which involves distinguishing an object

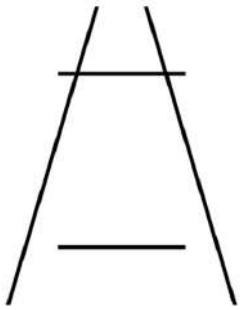


Figure 7: Illusion de Ponzo : Les deux lignes horizontales ont la même taille, mais celle en haut de l'image semble plus grande.

Part 2: Review of the Concept of Judgement

In this second part of the thesis, I will undertake a redefinition of the concept of judgement. Although Kant's theory, as I have suggested, serves as a useful starting point for critiquing artificial intelligence, it cannot be retained in its original form. The recent history of mathematics and physics seems to have dealt a severe blow to Kant's idea that we possess *a priori* structures, independent of experience. A striking example is found in geometry. Kant believed that geometric objects—and more generally, our representation of three-dimensional space—constituted an *a priori* form of human cognition: the intuition of space. According to him, mathematical truths were grounded in this intuition; that is to say, in the absence of such an intuition, it would be impossible to construct the theoretical framework of geometry.

Yet, since Kant's time, much water has passed under the bridge. Mathematicians such as Lobachevsky and Riemann demonstrated that it is possible to construct mathematical truths concerning non-Euclidean geometries. In these geometries, the sum of a triangle's angles does not equal 180 degrees. The notions of straight lines, parallels, or the "shortest path" are no longer entirely intuitive. It thus appears that, on the one hand, truth does not rest entirely upon *a priori* intuitions. On the other hand, it may be that our intuition itself is a mere construction. It is because we experience a three-dimensional world that we acquire a tridimensional intuition of space. But if we evolved in an aquatic environment, where objects deformed according to the currents, would our spatial intuition remain Euclidean? Would the "shortest path" between two points still be a straight line?

Taking these significant critiques into account, I will attempt to reconstruct the concept of judgement in such a way as to preserve the tension between analytic and synthetic judgements, yet without presupposing the existence of innate structures. I will argue for a circular interpretation of

---

placed against a visually noisy background.

the relationship between intuition and experience, which consists in stating that (1) structures are that without which perceptual information would not be possible; (2) but that these structures, in turn, are influenced by our perceptual experience. This constitutes a middle ground between a Kantian theory of human cognition and an empiricist theory.

Furthermore, I will suggest that human intelligence operates on three levels of processing:

- **Rationality:** This level comprises *rational concepts*, defined through rigorous and verifiable criteria. For instance, the idea of a "square" is a rational concept; it is precisely defined as “a plane figure with four right angles.” By contrast, "justice", a "house", or a "human being" are not rational concepts, for their definitions suffer from a certain vagueness. It is impossible to define with exactitude what a human being is, using words alone. One might say, of course, that it is “a biped endowed with such-and-such properties (language, reason, a certain genetic code...)”, but if the definition is too narrow, some humans might fail to meet it; while if it is too broad, non-human mammals could mistakenly be included. Rational concepts, then, are confined to the ideal realm of mathematics and logic. I will argue that it is possible for a computer to grasp rational concepts and manipulate them using top-down programming.

- **Sensation:** This level encompasses *sensible percepts*, defined as groupings of objects that exhibit similarities with one another. These percepts are not fully-fledged concepts, endowed with a definition. They are merely patterns of analogies, formed most often in an automatic and unconscious manner. A good example is the contemplation of an abstract painting. When you observe it, you perform a kind of segmentation by distinguishing certain colour forms from others. These forms are not governed by strict definitions. You cannot say that it is “a house”, a “human being”, or even “the colour red”. These forms exist only as contrasts; they are, therefore, the fruit of sensible experience, stripped of any rigid rational structure. In this essay, I will argue that it is possible for a computer to construct and manipulate percepts through bottom-up programming.

- **Understanding:** This The third level, caught between the other two, includes *intuitive concepts*, such as “justice”, a “house”, or a “human being”. These are indeed concepts, since they can be defined beyond our immediate experience, although they have no rational essence. When I look upon a “house”, I know that the example before my eyes is not the only possible or imaginable house. It is merely a singular case. Although I know this, it would be difficult for me to explain precisely

> what a house is in the most general sense of the term. If someone were to ask me, "What is a house?", my first reflex would be to list various types of houses: "Bungalows, villas, huts, flats — all of these are houses." In doing so, I am not truly answering the question. But is it even possible to give an exact definition of a house? Most likely not. That is why the concept of a house is considered intuitive. One can only conceive of it through an archetype (a typical house), and yet one knows that the notion of a house cannot be reduced to this archetype. In my essay, I will argue that intuitive concepts are the connective tissue between the sensible level of experience and the abstract level of reason. I will suggest that there currently exists no programming paradigm capable of convincingly constructing or manipulating them. In others words : intuitive concepts may be the a important blind spot in artificial intelligence.

The tricameral theory of cognition that I advocate can be seen as an extension of Daniel Kahneman's System 1 and System 2, for which he was awarded the Nobel Prize in Economics in 2002. Kahneman emphasizes the existence, within the human mind, of two systems of information processing: (a) an automatic, fast, statistical system, akin to the functioning of neural networks, and (b) a conscious, slow, rational system, resembling top-down programming. These two systems engage different neural regions and operate on distinct principles, yet both are essential to human decision-making. In his book titled *Thinking, Fast and Slow*, Kahneman provides the following description[14]:

> *System 1 operates automatically and quickly, with little effort and without a sense of voluntary control. System 2 allocates attentional resources to demanding mental activities, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration.*

I will propose the existence of a third system, situated between the first two. This System 3, which philosophy has traditionally called *understanding*, is a process that is conscious yet not fully justifiable. It consists in thinking concepts in a way that is neither statistical nor rational, but intuitive. This intuition operates through the formation of a mental image, a schema. This image, though inspired by concrete representations, designates something beyond all possible representations. What do I mean by this? When you represent to yourself the idea of a triangle or a tree, you may visualize an isosceles triangle or the typical shape of an oak, yet this

[14]This quotation is translated from English by myself.

mental image you produce does not refer to any particular isosceles triangle or oak—it refers to the abstract idea of the triangle or the tree. It is a universal archetype, a representation that aims beyond representation itself.

I will hypothesize that the higher level of cognition (reason) and the lower level of cognition (sensation) can be imitated by current algorithms, but not the intermediate level (perception or understanding). The higher level, in fact, resembles the top-down programming I discussed in the introduction. It consists in representing information in the form of symbols (numbers, algebraic characters) and manipulating them according to certain logical rules in order to reach a rationally valid conclusion. This is precisely what Isaac Asimov's robots do when they determine, based on the laws of robotics, whether a behavior should be deemed moral.

The lower level of cognition, on the other hand, resembles bottom-up programming. It consists in starting from various data points gathered from experience and associating them in such a way as to construct concepts. The algorithms used for this purpose rely heavily on statistics: neural networks, after all, are nothing more than multivariate statistical regressions. This results in adaptive, flexible responses that are context-sensitive but also prone to biases. A clear example can be found in ChatGPT, which tends to generate the most statistically probable responses, even when these are inadequate. There is, in fact, an invisible army—users, for the most part, but not only—who provide feedback to the algorithm, iteratively correcting its deviations.

The intermediate level (perception), finally, operates holistically. That is to say, it applies the conceptual structures of reason to the data of experience. This process is neither top-down nor bottom-up but rather circular. This means that data only become visible through structures, and structures, in turn, exist only through their contact with data. As a result, perception is comparable neither to the statistically blind association performed by neural networks nor to the narrow sequentialization characteristic of reason; rather, it emerges from contrasts. It is by attending to differences and similarities that the human mind carves the perceptual whole into individual units. Perceptual objects are the product of a principle of differentiation, which uses concepts to separate sensory content according to certain patterns.

Although this intermediate level may seem comparable to the functioning of neural networks, I will attempt to show, through concrete examples, that it relies on two fundamental intuitions: the intuition of the whole and the intuition of the infinite. These intuitions, as I introduce them, replace the intuitions of space and time found in Kantian theory. I will show that the Whole and the Infinite play a structuring role in our spatiotemporal representation of phenomena, that they in some sense sustain them, yet without assuming that space and time necessarily present themselves in a Euclidean form. One could say, in brief, that these are the most fundamental structures

of our human perception of time and space. I characterize them briefly here:

- **The intuition of the whole:** This intuition allows us to grasp information globally rather than sequentially. When applies to the very experience of time, it enables us to understand that the present moment does not exist without reference to a past moment and a future moment. Only once the moment in question has passed can we, retrospectively, take it as an object of thought, so that the process of thinking unfolds in a continuous back-and-forth between "what I am thinking now" and "what I think I was thinking a moment ago". In other words, our perception of time is global.

- **The intuition of the infinite:** This intuition allows us to comprehend and give meaning to ambiguity by weaving a web of associations between our intuitive concepts. It is at work whenever we relate an object to a concept. When we say "this tree is a birch", we both associate and partially differentiate the tree before us from the infinity of possible and imaginable birches. The singular tree derives its intelligibility from the infinite set of possible concepts to which it relates. Conversely, the concept of birch derives its intelligibility from the infinite set of possible singular trees to which it applies. This is why we sometimes learn as much from saying "this tree is a birch" as from saying "birches are like this tree". This commutativity between subject and predicate, specific to human language, allows us to grasp the link between universal concepts (birch) and singular objects (this tree).

Figure 8 summarizes this tricameral theory of cognition, for which I advocate. We observe that the human mind, represented on the left side of this diagram, possesses three faculties: sensation, perception (or understanding), and reason. These originate from the Cartesian "I think", that primordial state of perception in which the experience of being something rather than nothing is given as a totality. In this state, there is no distinction between subject and object, between thought, sensation, lived experience, and the external world. Only what is sensed, thought, and experienced in the present moment exists. However, the human mind decomposes this unity of "the present moment" to make independent objects appear. These objects evolve over time; they change, they transform.

The decomposition of the totality of lived experience into individual objects does not happen all at once; it is a process that takes place through successive back-and-forth movements between the three faculties of Sensation (unconscious percepts), Perception (intuitive concepts), and Rationality (rational concepts). These levels support one another: objects exist through

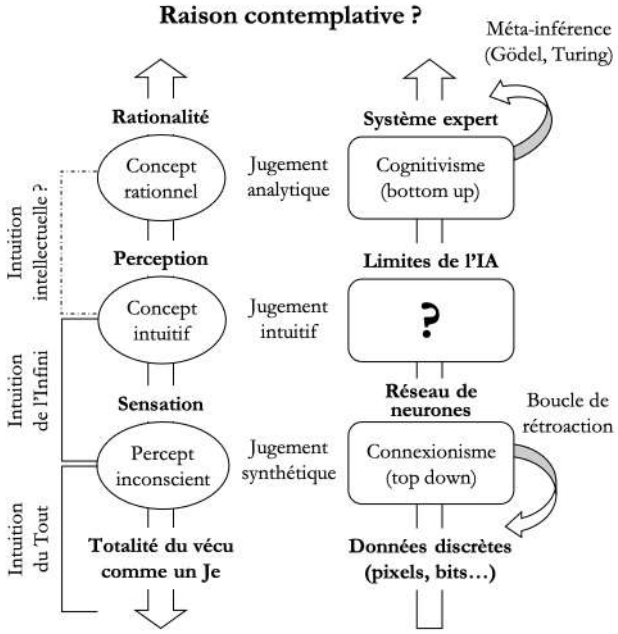


Figure 8

the whole, and the whole exists through objects. The transition from a lower level of cognition to a higher level—such as moving from a percept to a concept—is precisely what I call *judgement*.

There are thus three types of judgements. Those that transition from the totality of lived experience (the I think) to unconscious percepts (Sensation) are synthetic judgements and rely on the intuition of the Whole. Judgements that gather percepts (Sensation) to form intuitive concepts (Perception) are intuitive judgements and rely on the intuition of the Infinite. Finally, judgements that abstract the essential properties of intuitive concepts (Perception) in order to manipulate them in a logically deductive manner (Reason) are called analytical judgements and are based on a hypothetical intellectual intuition (affirmed by Plato but contested by Kant, for whom speculative Reason has lost all contact with the sensible world of experience).

The machine, represented on the right side of the diagram, functions differently from the human mind. It does not begin from an original *I think* but from a predefined set of independent discrete units (pixels, points, numbers). The task of algorithms is to associate these discrete units in order to reconstruct a global vision through a purely ascending movement (from parts to the whole). The machine, therefore, possesses only two levels of processing: (1) the level of neural networks, which in humans corresponds to Sensation (bottom-up programming paradigm), and (2) the level of expert systems, which corresponds to Rationality (top-down programming paradigm).

Part 3: The Intuitive Judgement in Front of the Machine

In this third part, we shall ask ourselves whether the intuition of infinity and the intuition of the whole— which I have described as the essential ingredients of human judgement— can be implemented through a Turing machine different from the ones we program nowadays. Indeed, the Turing machine constitutes the quintessential model of the computer. It consists of an infinitely long tape on which a reading head inscribes sequences of symbols such as 1 or 0. This head, nothing less than the sacred stylus of computer science, can read existing symbols, erase them, write new ones, and move as much as it pleases along the tape.

The question I will address in my thesis is whether a Turing machine would be capable of constructing sensible categories, such as the categories of a table, a book, or a birch tree. If the answer is affirmative, what would be the differences between these categories and those manipulated by the human mind?

I will suggest that the Turing machine encounters two theoretical obstacles.% What has been outlined above allows me to identify at least two major theoretical obstacles, each of which will deserve a thorough examination:

- **Commutativity of subject and predicate:** A Turing machine is defined by univocal assignments. When it stipulates $A = a$ by writing a given symbol on the reading tape, it means that the value a is assigned to the variable A. However, in the sensible categories of the human mind, we have seen that the relation of predication is commutative. The proposition “this is a tree” informs us just as much about the nature of the “this” standing before us as it does about what trees are in general. The equality relation $A = a$ is therefore not strict in sensible knowledge; it merely performs a pairing, a comparison, between the two terms. The categories we manipulate are not predefined classes that exist somewhere in the sky of ideas, waiting to be seized. On the contrary, they are a shifting process, evolving with our contact with sensible experience[15].

[15]This issue of subject-predicate commutativity explains why most of our concepts cannot receive a precise definition. It is also important to note that science, despite its rigor, does not escape this phenomenon. The notions employed in mathematics and physics are not defined univocally but illuminate one another. An example of particular importance for our discussion is the concept of "computability" in theoretical computer science. While we intuitively grasp that computability refers to the ability to determine, through an algorithm or a systematic reasoning procedure, the answer to a given question, it is legitimate to ask what exactly an "algorithm" or a "systematic reasoning procedure" is. Church-Turing's thesis states that these algorithms and reasoning procedures can be expressed through several formalisms of equivalent power: the Turing machine, Alonzo Church's lambda calculus, or the Herbrand-Gödel model of general recursive functions

- **The Idea of Infinity:** When constructing a percept as banal as that of a "glass" or a "tree," the human being manipulates the idea of infinity. Indeed, we are capable of synthesizing under a single mental image a countably infinite set of singular glasses or trees. Similarly, we can enumerate the numbers $1, 2, 3, ...$ until we asymptotically form the idea of a set of numbers $\mathbb{N}$ containing all the others. This idea of infinity is non-representational, meaning that we cannot explicitly represent it in a sensible manner, yet it is nonetheless manipulable by our minds. In other words, we can think of this set of numbers $\mathbb{N}$ without being able to enumerate its contents. The problem is that it is unclear whether a Turing machine could manipulate non-representational ideas, let alone the idea of infinity. While it does possess an infinitely long tape on which it can inscribe the numbers $1, 2, 3, ...$, can it form an idea of what lies at the end of the tape? Can it grasp that the finite symbol $\mathbb{N}$ designates an object infinite in its very contours?

In the framework of this thesis, I will hypothesize that each of these two problems—the commutativity of subject and predicate, and the one relating to the idea of infinity—corresponds respectively to the intuition of the whole and the intuition of infinity. These intuitions are the anvil on which the human mind, armed with the hammer of experience, forges categories. Without this anvil, experience would stir empty air. It requires resistance, a solid structure, to bring objects into being. Intuition, in its various facets, is thus the faculty of grasping objects—not through a traceable, logically deductive reasoning that proceeds linearly from premises to conclusions, but in a global manner. If we are to believe its Latin etymology, *intuition* derives from *intuitio*, meaning "the act of seeing at a glance." There is a tendency in philosophy to oppose the concept of intuition to that of reason. One need only think, for instance, of this statement by Schopenhauer: "Concepts form a special class of representations, entirely distinct from the intuitive representations discussed so far, for they exist only in the human mind[16]."

I will argue that the intuitions of the whole and of the infinite cannot be modeled using a classical Turing machine but require a transfinite Turing machine. The only element that differentiates the transfinite machine from

(see Appendix 3 for further details). Each of these models thus sheds light on the concept of computability; but not only does this illumination fail to exhaust the notion of computability entirely, it also requires a prior notion of computability to understand the relevance of these models. The human understanding of what is computable, far from adhering to a strict definition of the type $A = a$, unfolds within a network of interconnected concepts.

[16] Schopenhauer A., *The World as Will and Representation*, Chapter 9

the classical Turing machine is its ability to count infinity. Concretely, this means that the transfinite machine possesses an infinite reading tape and can reach the *"end of this tape"*. I will show that this extension of Turing machines is conceptually analogous to a machine capable of thinking the continuum rather than functioning in a discrete manner. For a continuous machine, a line is infinitely divisible, whereas for a classical machine, there exist atomic units of measurement beyond which no further calculation is possible. It is likely that the human mind can think the continuum, so much so that one is justified in believing that the transfinite Turing machine, in modeling continuous operations, captures a real aspect of human cognition.

Part 4: Towards a General Theory of Mind

The final part of this thesis will begin with the following observation: if the human mind possesses both the intuition of the infinite and the intuition of the whole, and if these intuitions can only be modeled through a transfinite Turing machine, how is it that our brain, by contrast, is finite in size? We must first acknowledge that the transfinite Turing machine serves as a model for intuition, but not as a model for the brain—whose mechanics are strictly finite. Furthermore, we must ask whether cognition truly originates from the brain. Or might we be mistaken in assuming that everything pertaining to the mind must ultimately be explainable by a physiological phenomenon occurring within our cranial vault?

In confronting the thorny question of the relationship between body and mind, I wish to defend a perspectival mysterianism. In philosophy, mysterianism consists in arguing that there are mysteries in nature that the human mind is not equipped to elucidate. The philosopher McGinn, for instance, maintains that consciousness belongs to this domain of mystery. I would like to follow in his footsteps, though with certain distinctions. I do not claim that consciousness, or mind as such, is inexplicable. I claim that the relationship between mind and body eludes all modeling. More precisely, the mysterianism I intend to advocate rests on three points:

1. The functioning of the brain can be modeled within a neurological framework.

2. The functioning of the human mind, understood from an introspective perspective, can be modeled within a psychological framework[17].

[17]This introspective model of the human mind would aim to explain the internal dynamics of the psyche from a subjective and phenomenological standpoint. Examples of such an approach can be found in Husserl's philosophy, which studies the principles by which objects are constituted in consciousness. One might also think of Kantian philosophy, which is not fundamentally far from a structural theory of the human mind, or even Freud's structural model, which seeks to make sense of concepts such as the Ego.

3. There exists no theory capable of bridging the biological model of the brain and the psychological model of the mind.

Objections immediately arise. Why, one might ask, should the link between body and mind resist all theorization? It is not so easy to assert, of a scientific or philosophical question, that it admits no solution. Indeed, no one can predict the future course of knowledge, nor how our very representation of a given problem will evolve. To take a historical example, it once seemed inconceivable to the ancients that the universe could have any shape other than that of a sphere, or that the Earth could be anywhere but at the centre of the known world. Yet despite these preconceptions, we now have theories in which the Earth lies on the outskirts of the galaxy, within an infinitely expanding universe whose structure defies Euclidean geometry. And this science, as advanced as it may seem today, remains in its infancy. Might not something as incomprehensible as consciousness, once freed from its biases, ultimately receive an explanation in the centuries to come?

The mysterianism I wish to defend holds that if the relation between body and mind is enigmatic, this is not due to a gap in our knowledge, but because the terms *"body"* and *"mind"* designate one and the same reality. This reality, however, we are incapable of perceiving in its unity. It appears to us as a cylinder would appear to a scientist living in a two-dimensional world. For such a being, the cylinder would manifest either as a circle or as a rectangle, depending on whether it was projected into the two-dimensional world through a longitudinal or transverse projection (see figure...). In other words, our two-dimensional being would see two distinct objects—a circle and a rectangle—and would have no *a priori* reason to believe they were two perspectives on the same three-dimensional object.

Similarly, we perceive reality sometimes through the prism of mind, sometimes through the prism of matter, and we suspect that there exists a link between these two perspectives—one that we seek desperately within the brain. We thus hypothesize (1) that mind and matter are two distinct realities and (2) that these realities converge somewhere in the brain. Furthermore, we assume (3) that the reality of mind must be entirely reducible to that of matter, such that all terms referring to mind (thoughts, ideas, sensations, pain...) are merely physiological phenomena.

I would like to suggest precisely the opposite: namely, (1) that mind and matter are one and the same reality, neither purely mental nor purely spatial, and (2) that reality cannot be fully explained either by mind alone or by matter alone. Therefore, it seems that our understanding of the world necessarily requires passing through both perspectives. Some problems receive an adequate response from the standpoint of mind, while others require

adopting the vantage point of matter. This amounts to a form of monism (only one reality exists) with a dual-aspect framework (this reality manifests in two distinct ways).

To establish the credibility of this philosophical stance, I will endeavour to show that dual-aspect monism is compatible with the practice of neuroscience. I do not intend to provide a sweeping overview of all influential neuroscientific theories concerning the representation of the human mind. Instead, I will focus specifically on the theory of consciousness developed by Gerald Edelman and Giulio Tononi. This choice reflects a dual commitment. First, it entails concentrating more specifically on the question of consciousness, and second, it aligns with a theory that lends itself to a metaphysical interpretation favoring dual-aspect monism. I will clarify and justify these two angles of approach as thoroughly as possible.

The first of my two commitments, the decision to study neuroscientific theories of consciousness, is particularly relevant when one recalls that the hypothesis of dual-aspect monism seeks to address the problem of articulating body and mind. What object of study, more than consciousness itself, confronts us head-on with this theoretical challenge? Consciousness cannot be subjected to neutral experimentation in the same way as objects that typically fall within the purview of scientific inquiry. It cannot be placed in front of a telescope or examined under a microscope, for it is the subject's inner life that is at stake. Though it lies at the heart of perception—or perhaps for that very reason—it appears to evade the known laws of diffraction. It is a blade that, though finely honed, cannot cut itself.

The second commitment consists in choosing specifically the theory of Edelman and Tononi, rather than, for instance, models of access consciousness inspired by William James and developed by neuroscientists such as Dehaene, Changeux, or Naccache[18]. This choice is justified by the particular methodology adopted by Tononi in his approach to consciousness. Rather than starting from existing theories of attention or cerebral functioning in order to isolate a set of key processes—such as those articulated around the thalamo-cortical system—he begins with the phenomenological experience of consciousness. In this experience, consciousness presents itself as the totality of reality, yet it obeys a certain number of organizational laws. He seeks to identify these laws through five axioms, after which he postulates the existence of a structural analogy between the functioning of

[18]These models of access consciousness are originally based on William James's intuition that there exists a higher nervous system that observes and controls the other cerebral systems. According to a lecture given by Stanislas Dehaene at the Collège de France in 2009–2010, these models rest on four ideas, formulated as early as the 1950s and 1960s and progressively gaining wider acceptance: (1) consciousness is associated with a central supervisory system, (2) this system has a limited capacity and must select which information to process (bottleneck system), (3) this system requires feedback loops, also known as reentrant processes, and (4) consciousness provides an internal space for synthesis, retention, and data sharing.

phenomenological consciousness and the physiological principles governing information processing in biological organisms. In other words, he aims to establish a neurological translation of principles first derived through an introspective approach.

This method of Tononi bears several similarities to my own thesis. First, it accepts the idea that a first-person perspective is inherently present in any scientific theory of consciousness. Rather than leaving this assumption implicit, it seeks to formulate its precise laws. Furthermore—and this is no minor advantage—Tononi's approach is compatible with double-aspect monism. Indeed, it does not reduce consciousness to material substrates but instead conceives of it as an information process. It is therefore permissible to interpret—though this is not explicitly his ontology—this information as a reality that can be understood through the complementary perspectives of mind and matter.

I will therefore endeavour to present Tononi's theory in detail, highlighting the influences it has received from Gerald Edelman. Since the two authors collaborated, it seemed difficult to separate them entirely. I will attempt to isolate the points of convergence and the divergences that gradually emerged in their respective works. My objective is not merely to show that neuroscience supports double-aspect monism, nor simply to argue that it could open new avenues for empirical intuition and research. Such an ambition would amount to a rather superficial discussion of the literature. I say superficial in the sense that it would not fundamentally justify the relevance of these authors within the framework of my thesis.

On the contrary, the analysis of Tononi will serve to lay the foundations—or rather, to establish the essential prerequisites—of a bi-perspectival theory of consciousness. This theory must take as its starting point the first-person perspective (introspection), but without assuming that this I-perspective is something given and immediate. Furthermore, it must grasp the articulation between the unity of "I think", "I feel", "I am", and the multiplicity of thoughts, sensations, and states of consciousness that constitute this unity. The mereological theory of consciousness[19] that I envision will be outlined toward the end of my thesis, drawing on the tricameral theory of the human mind, which distinguishes reason, perception, and sensation. It will be developed more extensively in a 300-page essay entitled Is Consciousness Dead?, which I have appended to the end of this document (see Annex 7).

[19]Mereology is the discipline that studies part-whole relations: the relations between a part and its whole, as well as the relations between several parts within the same whole (Cotnoir and Varzi 2021).

The Problem of Judgement in Philosophy

From analytical to synthetic

Chapter 1

Judgement in Plato: Origin of the Problem

We do not know, and we will not know.

Emil du Bois-Reymond

A question often debated in the history of ideas, one whose mere mention makes white-bearded scholars tremble, is whether philosophical problems are ever truly granted new answers. Observing how the edifice of philosophy has been constructed over the centuries, one might get the impression that every stone added to the summit of its rickety tower of metaphysical concepts was quarried from its very foundations—so much so that philosophy, rather than rising toward the heavens like physics or mathematics, has never ceased to stagnate.

If this is true, it would mean that our philosophical innovations are nothing more than a coat of paint, a layer of gilding, or a set of ingenious inscriptions applied to the cracked bricks of the thinkers who came before us. These embellishments might, admittedly, refine here and there the lines of inquiry opened by our predecessors, but their most important role would be to equip ancient philosophical positions with a contemporary language. This language consists in providing a logical structure, a mathematical formulation, to statements that had previously been nothing more than intuitions. Epistemological problems, which in Socrates' time presented themselves as disputes over definitions or universal principles, have, in our modern theories, been transformed into mathematical equations.

An example of this phenomenon—one that, I believe, few would dispute—can be found in the concept of atomism. The idea that the world is governed by the laws of nature, understood as the collision of atoms, is in essence an idea of Democritus. In ancient Greece, this idea was not necessarily dominant—far from it—but its implications had already been

examined in their most general, most abstract, and most conceptual terms. Among these, one might ask: Does the void exist? How is motion possible? What force moves matter? A sample of these reflections can be found in Democritus' argument below:

> *If every body is infinitely divisible, then one of two things must be true: either nothing will remain, or something will remain. In the first case, division cannot lead to pure and simple nothingness, for then matter would have only a virtual existence; in the second case, the question arises: what remains? The most logical answer is the existence of real, indivisible, and unbreakable elements—hence, atoms.*[1]

Modern science, as a paradigm, has had the merit of revisiting certain philosophical positions and granting them systematic development. Some practical problems—such as the ways in which atoms interact or the resultant properties of matter—have seen substantial progress, progress all the more impressive given that it has led to applications that have transformed the face of the world. The most fundamental enigmas, however—such as the relationship between void and matter, or, in another domain, the question of what existed before the Big Bang—remain largely unresolved. It would be an illusion to think that, simply because these problems have received a more precise mathematical formulation, we have made significant strides in understanding metaphysical mysteries.

Perhaps this deadlock stems from the fact that, while the scientific method is prodigiously effective at modeling the world, it does not allow us to answer the great philosophical questions? Modern knowledge, after all, follows Descartes' method, according to which a problem must be reduced to its simplest components, *"so as to ascend gradually, step by step, to the most complex knowledge"*[2]. But by what right can we be sure that the universe is decomposable into simple components? And if it is, is the decomposition unique, or is it akin to the work of puzzle-makers who, starting from the same image and following certain decomposition rules, produce a multitude of distinct puzzles? Might it not be conceivable that, by following a different methodological scheme, it is not matter, but rather the void, consciousness, or will that constitutes the foundation of all so-called "reality" (a word that, as Nabokov so aptly put it, may be the only one truly deserving of quotation marks)[3]?

Science does not share the same aim as philosophy. While philosophy begins with the simplest notions and gradually ascends toward conceptions

[1] This argument from Democritus is reported by Aristotle in *On Generation and Corruption*, Book I, Chap. 2, 9–19.

[2] Descartes, *Discourse on Method*, third rule.

[3] I refer here to a brief aside in the afterword to *Lolita*, which concerns a completely different subject.

"so paradoxical that no one wishes to believe them" (as Russell put it), science, by contrast, starts with the complexity of the world and forcefully reduces it to the simplest principles it can conceive. It never hesitates, along the way, to wield—and at times, to overuse—Occam's razor: *pluralitas non est ponenda sine necessitate* (plurality should not be posited without necessity). Its goal is not to produce complexity, but to cultivate confidence in the world.

For this reason, I will propose in this book that certain technical problems encountered in various scientific disciplines—including artificial intelligence—are, at their core, problems of concepts. By this, I mean that the difficulties involved are not *merely* a matter of empirical observation or of solving equations analytically, but that they also depend on the way we reduce a problem to its minimal units of analysis. Questions such as *"What is thought?"*, *"Can a machine judge truth and falsehood?"* will not be answered simply by increasing computational power or refining our understanding of the human brain. They also depend on the categories we use to describe the phenomena occurring in our consciousness. "Ideas," "sense data," "sensations," "perceptions," "memory," "reasoning"—all these are abstractions concerning a complex cognitive whole. The choice of concepts we manipulate, and the way we model them through our computers, bears an element of potential arbitrariness and reductionism.

Artificial intelligence methods are no exception to this. At a time when neural networks seem closer than ever to surpassing or imitating human beings in a variety of cognitive tasks once deemed impossible—including image creation, musical composition, and language use—it may be time to ask: What reductions have been made in the process? What has been lost of human consciousness's richness by forcefully feeding it into the data-churning maw of contemporary algorithms? Do machines rely on the same concepts as human judgement? And what, ultimately, can we expect from artificial intelligence?

Hypothèse 2 (Research Hypothesis)**.** *The obstacles facing the development of artificial intelligence are not merely technical difficulties; they also conceal epistemological problems, present since the dawn of philosophy, and linked to the very concepts we employ.*

One of the most crucial conceptual problems for artificial intelligence is precisely understanding how ideas arise in the human mind. We all experience seeing, in the space around us, a world populated by objects: a table stretches before us, standing on four legs, laden with books, glasses, and triangular biscuits. Each of these objects can be abstracted in our consciousness, detached, as it were, from the context in which we first perceived them. Furthermore, we can imagine these objects even in their absence, representing in our minds sometimes a shape, sometimes a colour, sometimes a general sensation.

However, it is not easy to understand how this abstraction—what we call an idea—comes to be. By what right does the vision of a triangle, for instance, relate to the concept of a triangle? The concept of a triangle is not any particular triangle, yet it encompasses all possible and imaginable triangles. Whenever I picture a triangle, mustering all my imaginative strength, what I see is not the concept of a triangle, but some particular triangle, endowed with a specific shape, certain angles, a colour, even, as it were, a scent. Without the concept of a triangle—an entity that seems to elude all representation—would I be able to recognize triangles in nature? To isolate them? To manipulate them in thought? To discover properties regarding their angles?

In the following chapters, I would like to show how these problems appear under Plato's pen. Choosing to speak to you about the father of Greek philosophy might betray a lack of originality. I would not want to come across as one of those rambling scholars who, in place of thought, merely cling like ivy to the trunk of ancient thinkers. However, I have two reasons for discussing the Socratic dialogues.

The first, and by no means the least, is that I wish to demonstrate that the epistemological problem faced by artificial intelligence is not a technical issue. This assertion means that the problem can be articulated without resorting to the formalism of contemporary mathematics and that it was already present at the very origins of philosophy. All truly fundamental problems—those that pertain to the very nature of thought—were addressed by the first human beings capable of objectifying thought and setting it down in writing. And the oldest preserved philosophical writings that have reached us come precisely from ancient Greece.

The second reason, even more significant, is that Plato constructs concepts that are not without relevance for clarifying certain difficulties encountered in contemporary artificial intelligence. One of the merits of the master of the Academy in Athens lies in his distinction, within human thought, between three kinds of judgement: experiential judgements, true judgements, and true judgements accompanied by justification.

Experiential judgements, first of all, can be seen as a certain capacity for synthesis; they relate to the real world by means of observation, comparison, and imitation. True judgements accompanied by justification, on the other hand, rest on an aptitude for logical reasoning. They begin with certain principles and ascend toward more general conclusions. True judgements without justification, finally, occupy an intermediary position. I would like to show that, in philosophy, they may represent the most archaic prototype of that specifically intuitive form of judgement I mentioned in the introduction (see 1). These judgements are neither purely empirical nor purely logical. They establish a link between sensation and reason, between experience and deduction.

1.1 Socrates vs. Protagoras: Monologues of the Soul

To clarify the precise role that Plato assigns to judgement, reading a text entitled the *Theaetetus* offers several points of interest[4]. First, in terms of its method, this foundational text of Plato introduces us to the use of *logos* in philosophical inquiry; in terms of its object, it raises the question of what makes a science of change possible; finally, in terms of its polemical aim, it is in close dialogue with the sophists and the pre-Socratics. This is why I wish to rely on this work to suggest how Plato breaks with the earlier tradition—that of the sophists—and, in doing so, leads philosophy toward a specific kind of judgement: the synthetic *noetic* judgement.

The *Theaetetus*, like many Socratic dialogues, seeks to uncover, through an exchange between several characters, the essence of a concept. The concept in question here is science, understood as true knowledge. It is easy to see why this theme is of particular importance to our inquiry, as it concerns the question of how the human mind can discern the difference between truth and falsehood.

It all begins as Socrates, visiting his friend Theodorus, asks the latter to introduce him to the finest and most talented of his disciples. In response, Theodorus presents him to the young Theaetetus, whose beauty lies in his soul. He paints the portrait of a perfect philosopher: "I have never seen anyone so admirably gifted. That a man should have a facility for learning so exceptional that one would be hard-pressed to find its equal, that he should be supremely gentle, and that he should possess, in addition to all this, an unparalleled courage—I would never have thought it possible to encounter such a combination, and I know of no other example[5]." Socrates, delighted, begins to question the young man. He asks whether science is wisdom or whether there exists a particular distinction between these two types of virtues.

As one might expect, the young man's initial answers to Socrates are unsatisfactory. When asked what science is, he lists its various disciplines: geometry, astronomy, arithmetic, and music, he says, are sciences, but so too are shoemaking and the various crafts of artisans. In doing so, however, he fails to reveal its essence. Socrates, therefore, has little trouble correcting him: "You are not defining science itself, but its objects. If I were to ask you what mud is and you answered, 'There is the mud of potters, the mud of brickmakers, and others,' I would be no wiser as to the nature of mud. What you should have told me is that mud is earth mixed with water[6]."

[4] In the remainder of the text, I will cite passages from the *Theaetetus* using the Stephanus pagination system. The French translation used is that of Chambry.

[5] *Theaetetus*, 144a.

[6] cf. introduction to the *Theaetetus* by Chambry, p. 44.

Socrates immediately clarifies that he does not possess the answer to their inquiry, but he is capable of assessing the original ideas Theaetetus will propose. Moreover, he is able to stimulate his intelligence by examining with him the theories of those who have previously addressed the question. In this respect, he operates somewhat like a midwife of souls: "Have you not heard that I am the son of a midwife, and that I have the art of delivering minds, just as she had the art of delivering women? Furthermore, I know how to discern whether a young man's mind is bringing forth a mere illusion or a true and real fruit. Trust me, then, and do not be disheartened if, in examining what you say, I judge it to be but a phantom without substance[7]."

They will then consider, one after another, three definitions. The first, which is also the one on which Plato dwells the longest, holds that "science is nothing other than sensation[8]." This is, in fact, Protagoras's view. The second, more refined, asserts that "science does not reside in impressions, but in reasoning about impressions[9]," which Theaetetus expresses by saying that science is true opinion. The third, finally, is an adaptation of the previous one, maintaining that science is true opinion accompanied by reason—that is, furnished with a set of more analytical definitions of the terms involved in the opinion. Each of these three approaches will face its own set of aporias. However, their analysis provides Plato with the opportunity to expound upon and critique the prevailing theories of his time, particularly that of Protagoras.

1.1.1 Judgement of Experience

There is a reason why Plato places so much emphasis on the first thesis—that of sensation. It seems to have been the position of Protagoras, one of his most prominent rivals. Protagoras' thesis is the following: knowledge is sensation, and sensation is ever-changing. By way of illustration, he offers the example of hot and cold. The same wind that causes shivers in one person may feel mild to another. So what must be held as true is not some kingdom of objects to which we gain access through the senses, but rather the subjective—and entirely relative—sensations that we experience. The judgement of science is thus conceived as a *judgement of experience*.

Plato raises a series of objections to the judgement of experience. The most famous consists in accepting Protagoras' thesis—that truth is entirely relative to the point of view adopted—and showing that it leads to absurdity. It works as follows: suppose that Protagoras holds true the proposition P, according to which "Man is the measure of all things." At the same time, suppose that Socrates denies this proposition. From Protagoras' own standpoint, the proposition must be both true (since things are measured by

[7] *Ibidem*, p. 45.
[8] *Theaetetus*, 151e.
[9] *Theaetetus*, 186d.

Protagoras) and false (since they are equally measured by Socrates). The formalization of this reasoning is not without resemblance to a well-known syllogism:

Protagoras believes that Man is the measure of all things.
But Socrates is a Man.

Therefore, Protagoras believes that Socrates is the measure of all things.

Yet since Socrates holds P to be false, we arrive at the paradoxical conclusion that Protagoras does not believe that Man is the measure of all things. Has our good sophist been caught in his own trap? Socrates concludes, with irony: "Suppose that he [Protagoras] himself did not believe that Man is the measure of all things, and that the majority do not believe it either—as indeed they do not—would it not then follow, by necessity, that the truth as he has defined it exists for no one[10]?"

1.1.2 True Judgement without justification

After harshly contesting sensualism, Socrates eventually convinces Theaetetus to propose a new definition. If knowledge does not originate from sensory impressions, it must arise from reasoning about those impressions—that is, from *judgement*. The introduction of this term, I believe, deserves emphasis, for it marks a decisive shift in the history of thought. Alongside the capacity to feel, it posits another kind of faculty. But what does "to judge" mean in the mouth of Socrates? It is "the name [...] given to the soul when it applies itself alone to the study of beings[11]." According to Socrates, judgement is thus the soul's ability to engage in a dialogue with itself, in a mode of inner discourse, in order to uncover the beings to which sensations correspond. It is likely that the type of dialogue in which the soul engages is analogous to the one unfolding between Socrates and Theaetetus.

Of course, knowledge does not consist indiscriminately of all possible judgements, but only of those that are true. Theaetetus thus advances the following suggestion: "To say that every opinion is knowledge, Socrates, is impossible, since there is also false opinion. But it is likely that true opinion is knowledge, and that is my answer[12]." Socrates then asks him to explain what makes an opinion false. He goes even further: he asks what makes false judgements possible. In other words, he requests that Theaetetus provide a model of knowledge that can account for both error and ignorance.

Socrates then undertakes, alongside Theaetetus, to sketch the contours of a theory of knowledge. To do so, he resorts to an allegory—that of the wax block: "Suppose, for the sake of argument, that there is in our souls

[10] *Theaetetus*, 170e–171a.
[11] *Theaetetus*, 187a.
[12] *Theaetetus*, 187b.

a block of wax[13]." Whenever the mind wishes to remember something it has seen, heard, or conceived, the imprint of that thing is pressed upon the wax. Under this hypothesis, remembering consists in retrieving the imprint within our soul and matching it with the appropriate sensible object. If, for example, one bears the imprint of Theodorus and Theaetetus, one must recover this memory within the soul and correctly align it with the persons standing before one's eyes. But when the imprint has not been made or has faded away, it becomes impossible to recollect the thing correctly. Error, then, lies in a faulty correspondence between imprints and sensible objects.

As convincing as this theory may seem at first glance, it does not fail to arouse Socrates' sharp skepticism. The allegory of the wax block appears to suggest that knowledge is always a matter of correspondence between judgement and sensation. Yet, is it not possible to reason within the intelligible realm? Is error not also possible in the absence of sensible objects? One need only think of mathematical demonstrations to realize that the notions of truth and falsehood cannot, unless one adopts a strictly empirical conception of mathematics, be reduced to a mere correspondence between thought and sensation. The example he gives is that of mental calculation. Has it never happened that a man, contemplating the numbers 7 and 5, comes to believe that their sum is 11? "We must therefore show that false opinion is something altogether different from a discord between thought and sensation. If it were merely that, we would never err in our pure thoughts[14]."

Socrates then sets out to clarify the meaning of the terms "to know" and "to possess knowledge." To begin, he draws a distinction between "having knowledge" and "possessing" it. Just as one may possess a cloak without presently wearing it, one may be the holder of knowledge without presently having it in mind. As soon as thought is engaged, as soon as memory is stirred, the concrete ideas corresponding to this knowledge resurface in the soul. It is as if each piece of knowledge possessed were a bird in an aviary. When the mind wishes to summon this knowledge and use it in reasoning, it must seek out the bird within its enclosure, which requires a certain effort. The task, to be sure, is easier than capturing a wild bird; applying knowledge is simpler than learning it for the first time. Nevertheless, the possibility of error remains. There is always the risk of mistaking one bird for another, or of failing to catch the one corresponding to the knowledge sought.

This explains two things: first, the possibility of errors in reasoning, such as miscalculations; second, the fact that even a skilled arithmetician, who knows all numbers, may still seek to discover their properties. Possessing knowledge of numbers does not mean having the numbers, along with their characteristics, presently before one's eyes. We shall return to this theory in greater detail when we compare it to Kant's conception.

[13] *Theaetetus*, 191c.
[14] *Theaetetus*, 196c.

1.1.3 True Judgement with Reason

A doubt always lingers, insidiously creeping into even the most resolute mind. Shall we call a man learned when a true opinion slips accidentally from his lips? Or do we expect him to be able to justify his belief? Socrates takes the example of a judge who, in order to render a verdict, relies on a set of eyewitness accounts. He does not personally know the situation, for he neither lived it nor witnessed it himself. By his own admission, knowledge is not on his side, but on that of the victim, the criminal, and those who were present. Yet, he may still reach the right decision. "If true opinion in judges and knowledge were the same thing, then the best judge would never hold a correct opinion without knowledge. But in reality, it appears that they are two different things[15]."

This leads Theaetetus to propose a third and final definition: "True opinion accompanied by reason is knowledge, but [...] without reason, it lies outside the domain of knowledge[16]." But what exactly should we understand here by "reason" (*logos*)? Definitions? Arguments? Socrates expounds a theory. The mind, he says, can know nothing of the fundamental elements that constitute all reality. At most, it can name them, but not describe them: "Each element taken on its own can only be named, and nothing more can be said about it—not that it is, nor that it is not ... nothing must be added to it, if one wishes to express this element alone. One should not even attach to it the words 'the,' or 'this,' or 'each,' or 'alone,' or 'this one,' nor many other similar terms[17]." By slightly extrapolating Socrates' words, it may not be too bold to suggest the following: the knowledge of an object consists in linking the soul's correct intuition (which names the object and belongs to the synthetic order) to a correct rational explanation (which analyzes the object by breaking it down into its elements and belongs to the analytic order).

I wish to draw attention to the fact that the structure of ideas in the *Theaetetus* is not left to chance but rather reflects a certain anthropology. Plato begins with what pertains to sensation, which is characterized by change. The question he poses is that of the relationship between the sensing organ and the sensible object. Over the course of the dialogue, he refines the issue: sounds and colours are perceived through the ears and eyes, but are they perceived *by* them or *with* them? This nuance is crucial. If sensible objects are perceived *with* the organs, this means that the subject is endowed with sensory faculties that are grafted onto the body and that allow him to perceive them. But if they are perceived *by* the organs, this implies that it is a single faculty of the soul that operates through the organs and that, by means of them, perceives sensible objects. These two approaches are

[15] *Theaetetus*, 164b.
[16] *Theaetetus*, 201c-201d.
[17] *Theaetetus*, 202a.

radically different. In the first case, the senses are self-sufficient, and there is no need to introduce another order of intelligibility. In the second, they are merely instruments of a higher faculty.

To prove the existence of this faculty, Plato sets out to show that certain ideas—those concerning "being" and "non-being"—cannot originate from the senses. Consider a sound and a colour. It is impossible to know colour through hearing, just as it is impossible to know sound through sight. And yet, one can form ideas about both sound and colour that relate equally to one and the other. This is the case, for instance, with the idea that both sound and colour *exist*, or with the idea of their similarities and differences. This applies, in short, to everything that concerns their being. Socrates then asks: "Through which organ do you conceive all this about these two senses? It is neither through hearing nor through sight that one can grasp what they have in common[18]." There must, therefore, exist a faculty that transcends the senses. This faculty is the one that makes use of language. It depends on no bodily organ but is exercised by the soul itself. It alone is concerned "with being and non-being, with resemblance and dissimilarity, with identity and difference, and also with unity and the other numbers applied to these things[19]."

Thus, a distinction must be made between sensory impressions—such as "hard," "soft," "cold," and "hot"—which are given by bodily organs, and the ideas formed about these impressions, which are judgements of the soul. Only the latter pertain to the being of things and their truth. "There are, then, certain things that men and beasts are naturally capable of sensing from birth: these are the impressions that reach the soul through the body. On the other hand, the reasonings made about these impressions—regarding their essence and their utility—only come with difficulty and over time, through effort and study, to those in whom they take form[20]."

Far from stopping at the faculty of "sensation," which allows us to perceive the world through the senses, Plato insists on the existence of a faculty of the soul that is purely intelligible. He calls this faculty *reason* (*logos*). This term is introduced for the first time when Socrates and Theaetetus realize that judging is not merely a matter of matching ideas with impressions but that there also exists a form of pure thought, unrelated to any sensory content. Such an operation is exemplified by mental calculation. "We must therefore show that false opinion is something altogether different from a discord between thought and sensation. If it were merely that, we would never err in our pure thoughts. But in reality, either false opinion does not exist, or it is impossible not to know what one knows[21]." As long as ideas relate to sensible things, truth and falsehood arise from their correct corre-

[18] *Theaetetus*, 185b.
[19] *Theaetetus*, 185b-185c.
[20] *Theaetetus*, 186b-186c.
[21] *Theaetetus*, 196c.

spondence. But once one ascends into pure thought, error—if it exists—can stem only from the proper or improper use of reason.

The precise nature of reason as a faculty, however, is not thematized in the *Theaetetus*. On the contrary, in many respects, it appears to be nothing more than an operation of language. One has the impression that the dialogue's conclusion struggles to clarify the concept of reason. The definitions considered, only to be successively refuted, are likely a compilation of ideas defended by Plato's contemporaries, such as Antisthenes or the Megarian school. Whatever their origins, the multiplicity of these theories attests to the importance of this question at the time in philosophical circles and, at the same time, reinforces the inaugural nature of Plato's response when he chooses to make the nascent concept of Reason the cornerstone of all knowledge. It remains to be determined precisely how he defines this faculty and what place he assigns it within his theory of judgement.

1.2 The Contemplation of Beauty

I have indicated that Plato envisions three types of judgement: empirical judgement, true judgement without justification, and true judgement endowed with reason. The first is based on observation of the world and represents the properly synthetic side of knowledge. The last is grounded in reasoning and, in contrast, represents its analytic side. Between these two ideal cliffs—shaped by ambitions and shattered dreams—stands a judgement that is true but unjustified. Yet Plato does not attribute to this judgement any greater value than to a mere empirical judgement. He does not entertain the existence of a middle path between synthesis and analysis. In Socrates' words, there is no excluded third: either something belongs to the senses—and is therefore potentially deceptive—or it belongs to reasoning.

This dichotomy between the senses and reason may well be foundational to the Western conception of rationality. Is not the philosophical vocation of the ancient Greeks, and later throughout the Middle Ages, characterized by a generalized suspicion toward the senses? With the exception of the English empiricists and the early atomists, truth is most often placed on the side of analysis. This bias disqualifies judgements that are neither justifiable nor arguable. One can understand why such judgements would struggle to find a place within a rational collective life, where truth is subject to the exchange of viewpoints, debate, and deliberation. It is undoubtedly difficult to separate the notion of truth from a broader social and political democratic project.

Be that as it may, this binary vision of judgement, as found in Plato, is accompanied by a bicameral representation of the human mind. The soul, in Socrates' account, possesses two faculties: sensibility and reason. The first concerns itself with empirical judgements; the second, with logical

analysis. Each of these faculties, in turn, is divided into corresponding sub-faculties. Sensibility contains a dimension of imagination (*ekasia*) and belief (*pistis*), while reason comprises a calculating dimension (*dianoia*) and a contemplative one (*noēsis*). This structure of human knowledge finds a striking illustration in one of Plato's most important works: the *Republic*. I would like to briefly outline this Platonic theory, as it will prove useful in the continuation of my thesis.

1.2.1 Analogy of Plato's Divided Line

Book VI of the *Republic*—perhaps one of the most important in Plato's work—opens with a dialogue between Socrates and Glaucon on the nature of the true philosopher and his role in the city. At a certain point, weary of hearing him refute the theories of others, Glaucon and several other listeners implore Socrates to speak of his own views. They ask him to reveal to them the nature of the Good, to offer them a definition at least as clear as the one he had recently provided on the question of justice.

Faced with such demanding expectations, Socrates warns that he will not be able to speak the Good itself, for it belongs to the realm of the intelligible. However, he can attempt to present one of its sensible metaphors: the sun. In what sense, exactly, does he claim that the Sun should be likened to the Good? In the sense that light is the indispensable medium of vision. Let us explain. When the eye turns toward objects illuminated by the sun, it can see them clearly. Likewise, when the soul turns toward the Ideas illuminated by the Good, they appear to it in their full extent. «Just as he is [The Good], in the intelligible realm, to intellect and the intelligibles, so this [the Sun] is, in the visible realm, to sight and the visibles. [...] When she [the soul] turns toward that which truth and being illuminate, then it thinks, it knows, and it seems to possess intellect[22].»

Socrates pursues the analogy with greater specificity. The sun, he explains, is the cause of both light and sight, just as the Idea of the Good is the cause of both truth and knowledge. Thus, knowledge and truth are products of the Good and, as such, must be part of its definition. However, the Good cannot be reduced to them. By its very nature, it is something more precious than any Idea one might conceive of it. And with good reason: it is the origin of all Forms. Socrates states that the Good «is beyond being (*epekeina tēs ousias*)[23]», thereby completing his description of its perfect transcendence.

Buoyed by his success, Socrates proceeds to draw a line in the sand and divide it into four segments of equal length. The first two segments, he explains, represent the two levels of reality within the intelligible world: that of mathematical reasoning and that of the pure Ideas of the Good, the

[22] *Republic*, 508b-508c.
[23] *Republic*, 509a.

Beautiful, and the True. The last two segments, by contrast, designate the two levels of reality within the sensible world: that of deceptive imagination and that of empirically grounded beliefs. The knowledge we have of the sensible world falls under opinion (*doxa*), while the knowledge we have of the intelligible world depends entirely on science (*epistēmē*). Opinions are fickle and changeable, shifting with the tides of our contradictions. Science, on the other hand, is detached from experience and absorbed in the contemplation of Ideas, in a way that both clarifies and expands upon the vision set forth in the *Theaetetus*.

1.2.2 The Pathemata

Each segment of the line is associated with a *pathēma*, an «affection». This Greek word, derived from the verb *paschein*, means "to undergo" or "to suffer," in contrast to *poiein* ("to make, to act") or *prattein* ("to accomplish, to act")[24]. In the present context, it denotes a subject's disposition to experience something upon contact with a sensible or intelligible object. The four *pathemata* thus define as many «faculties»—whether sensible or intellectual. Accordingly, there exist two faculties, reason and sensation, each of which is further subdivided into two sub-faculties.

Noèton *Epistémé* (science)		**Horatón** *Doxa* (opinion)	
Noèsis (intuitive)	*Diánoia* (discursive)	*Pistis* (belief)	*Eikasia* (imagination)
↓	↓	↓	↓
concerns the anhypothetical	concerns the hypothetical	concerns sensible objects	concerns their images

Eikasia is a term rarely found in Plato's texts. In the entire *Perseus* text corpus,[25] it appears only five times, three of which occur in the *Republic* alone. To better grasp its usage, it is useful to refer to the verb *eikazein* ("to represent, to compare, to conjecture"), of which it is the action noun and which appears far more frequently. Commentators have variously translated *Eikasia* as "simulation," "imagination," "illusion," "the faculty of reasoning based on resemblances," or "representation." Pachet's definition seems to me the most relevant in the present context, for what Plato appears to have in mind is a capacity to judge sensible things based on appearances. It is, therefore, the most "debased" form of knowledge.

[24] cf. annotations on the *Republic* by Suzanne [**?**], note 17.

[25] *Perseus* is a digital library project initiated and maintained by Tufts University: http://www.perseus.tufts.edu/

Pistis is translated as "faith," "belief," "conviction," or "assurance." It expresses the idea of a certain attitude of trust toward sensory testimony—not only sight but also hearing, touch, smell, and so on. It is the disposition of believing that a sensible reality stands before us and that it is as we represent it to be. In this sense, it may not be too bold to compare it to what Husserl would later call the "natural attitude." Yet, far from stopping at first impressions, this attitude seeks verification and objectivity by questioning sensory data through experience. It is, in essence, the spirit of "science" in its most empirical and practical stage, reminiscent of Lachès' assertions about courage. In more Kantian terms, we might say that this form of *doxa* consists of synthetic *a posteriori* judgements.

Dianoia has been translated as "discursive knowledge," "discursion," "reason," or simply "thought." Its meaning largely overlaps with that of *noēsis*, which we will examine below, but it differs by the addition of the particle "*dia*." What is the significance of this prefix? In attested usages, it can convey (1) separation, in the sense of distribution or division; (2) penetration; (3) superiority; or (4) completion.[26] The most relevant meanings here are likely those of separation or division.

The faculty Socrates has in mind concerning this segment is properly discursive. It consists in structuring the content of sensibility in such a way as to extract intelligible principles from it. Just as a geometer decomposes geometric figures into points and segments, one must first divide sensible images or words into as many analytical elements as necessary. By positing a number of hypotheses, one then ascends toward certain principles or theses of universal validity. The archetype of this reasoning is that of mathematics in its formal and apodictic dimension. One might think, for instance, of Euclid's *ad geometrico* demonstrations concerning the fact that the sum of the angles in a triangle equals 180 degrees. The associated discourse is logical-deductive.

To complete this definition, it should be noted that the prefix "*dia*" in *dianoia* may also refer to the fact that *dianoia* occupies an intermediary position on the line of being. On one side lies pure intelligibility, while on the other stand images. *Dianoia* can thus be conceived as the faculty mediating between knowledge of sensible images and knowledge of Forms. Moreover, it operates through a *dia*-logue that the soul conducts with itself.

Noēsis, on the other hand, has been translated as "intelligence," "intellection," or "spiritual activity." It is formed from the verb *noein* by means of the suffix *-sis*.[27] Less commonly used than the Greek word *nous*, which frequently designates the mind or intellect, *noēsis* has the advantage of referring specifically to intelligence as a faculty, rather than in its other meanings. What is at stake here is thus a particular capacity to apprehend the

[26] *Ibidem.*
[27] *Ibidem.*

intelligible.

Unlike *dianoia*, which is discursive, *noēsis* does not consist in starting from hypotheses and proceeding by deduction. Instead, it questions the very status of these hypotheses—both epistemologically (why are they true?) and ontologically (do they correspond to a reality existing beyond our knowledge of them?). It is a quest for the first, unhypothetical principle that allows us to conceive the true nature of things.

This first principle is the Good. It does not fall under ordinary discursive reasoning, which is why *noēsis* can only be of an intuitive nature. It is, properly speaking, that part of reason which, when freed from the senses, may be led to the revelation and contemplation of the Idea of the Good, the Beautiful, and the True—an inseparable triad. For this reason, we shall translate it as "intuitive reason" or even "contemplative reason." Socrates says of those who employ it that «they make use of hypotheses, not as principles, but as what they truly are: stepping stones and supports from which to ascend to what is unhypothetical, to the principle of all things[28].» As we shall see in the next chapter, this idea closely resembles what Kant calls Pure Reason, but in a more contemplative sense—one that is, in some respects, almost mystical. Modern philosophy has stripped Plato's thought of its poetic dimension while retaining the notion of "intuition" or "self-evidence."

In conclusion, I would argue that Plato's essential move, in the matter at hand, was to break with Protagoras' assumption that truth and the verdict of the sensory organs are identical. Once language ceases to be a mere sensation, it must, alongside reason, seek the being of things in their purely intelligible dimension. This search passes through impressions, examines them, but is ultimately realized in judgement, whose task is to articulate their unity, their universality—in other words, their very being. In the highest vision of the world, which is that of intellectual contemplation, there is no distinction between the Ideas of the Beautiful, the Good, and the True. All forms of nature merge. They echo one another, complement each other, in perfect unity. This is why, for Plato, the highest form of aesthetic experience is also a moral and scientific experience. Behind the world, there is something universal that erases the boundaries perceived by the senses between objects. When the subject strives toward this unity, they undergo an inner transformation. They are purified. They become fully human.

[28] *Republic*, 511b.

What to take away from this chapter

From the very beginnings of philosophy, Plato distinguishes three kinds of judgement:

1. *Experiential judgement* is synthetic. It operates through the faculty of Sensation and consists in observing the sensible world.

2. *Rational judgement* is analytic. It operates through the faculty of Reason and consists in discovering the essence of things—their definitions, we might say.

3. *Unjustified true judgement*, allows one to subsume different sensible realities under a single concept but without possessing an explicit criterion for doing so. Unlike the other forms of judgement, Plato does not associate this one with any particular faculty. He merely states that it constitutes the very exercise of the human soul.

As we shall see in the next chapter, it is upon this intermediate unjustified true judgement that Kant will build his theory of human understanding.

Chapter 2

Judgement in Kant : Genesis of the Understanding

I showed in the previous section that Plato identifies four types of cognitive aptitudes: *eikasia* (imagination), *pistis* (belief), *diánóia* (discursive reasoning), and *noèsis* (intellection). The first two belong to the faculty of Sensation, while the latter two pertain to Reason. In the sensible world, judgements are established by means of the senses, experience, and measurement. In the intelligible realm, by contrast, truth is a matter of rational deduction—proceeding from first principles and gradually ascending toward higher truths.

To illustrate the difference between sensible judgement and rational judgement, Plato gives the example of a geometer who draws a triangle on the ground and measures the sum of its angles to be 180°. As long as he relies on the sense of sight, he remains in the domain of the sensible. He observes patterns, looks for counterexamples, establishes correlations. His judgement may be true, but only by happenstance. The proper way to issue a judgement, according to Plato, is to offer a proof grounded in definitions. It is by starting from the concept of a triangle and the concept of an angle—by examining what is contained in their essence—that one arrives at the certain conclusion that a triangle satisfies this or that geometrical property.

One may say, without taking much risk, that Euclid's *Elements* constitute Plato's model of truth. The *Elements* illustrate the idea that a triangle, an angle, a circle, are of a different nature than the geometric shapes drawn in sand or on grainy paper. They possess an essence, a definition, from which necessary, demonstrable, and eternal properties flow.

It follows that two species of judgement may be discerned in Plato: those that are true without logical justification, because they are grounded in ex-

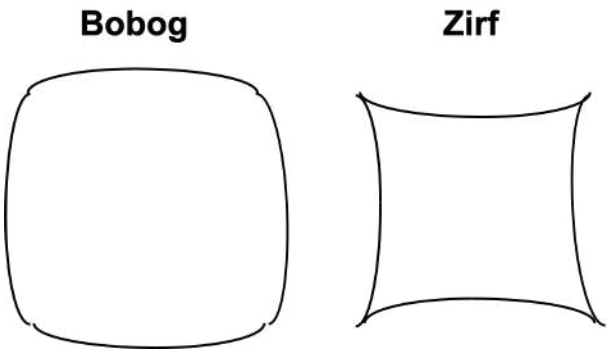


Figure 2.1: Two unknown geometric figures.

perience of the world; and those that are true and accompanied by a genuine explanation. These two types of judgement prefigure the later distinction between *synthetic* judgements, grounded in experience of the world (for example: "does this object resemble that one?"), and *analytic* judgements, based on reason alone (for example: mathematical demonstrations).

There is, however, a middle path, which appears in Plato's text, even if it is not explicitly thematized or clearly distinguished from the other two. This path concerns the true judgement without justification—a judgement which, while remaining within the domain of rational concept, nonetheless refers to sensible objects. This kind of judgement arises when the thinking subject attempts to connect not two ideas (the idea of a triangle and the idea of an angle), nor two sensible objects (the drawing of a triangle and that of an angle), but rather an idea with a sensible form. One might think, for instance, of someone who sees a figure drawn in the sand and exclaims, with a fervor worthy of Archimedes discovering the laws of hydraulics: "Look, it's a triangle!"

The kind of judgement I am referring to is not equivalent (1) to the rational or analytic judgement; nor is it equivalent (2) to a synthetic judgement based purely on the senses. To see this more clearly, I would like to present a small thought experiment of my own invention.

2.1 The Intuitive Abstraction

Suppose that you are presented with two unfamiliar geometric figures, bearing the names "Bobog" and "Zirf" (see figure 2.1). Based solely on this initial example, you are asked to sort a stack of drawings according to whether they are Bobogs or Zirfs, without being given any precise definition. To assist you, your colleague has annotated each figure with either the label "Bobog" or "Zirf," but you do not know whether these annotations are accurate or not (see figure 2.2).

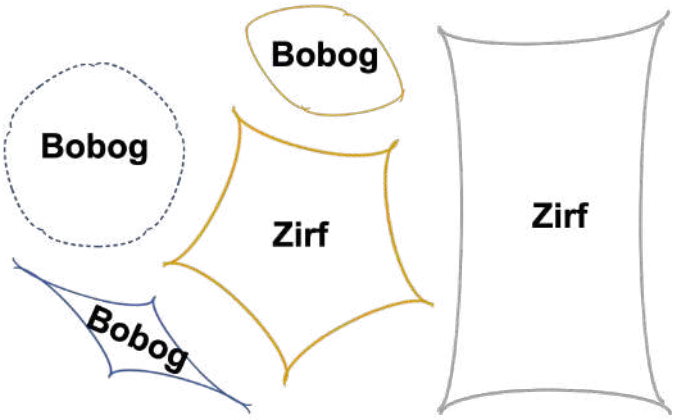


Figure 2.2: A collection of geometric figures annotated by a colleague. Are they Bobogs or Zirfs? Your colleague's annotations may not be reliable.

The first strategy consists in going through each image one by one, paying close attention to the annotations. If the word "Zirf" appears, you place the drawing in the left pile. If the word "Bobog" appears, you place it in the right pile. This approach is entirely rational. For this method, there is no need to actually identify the shapes depicted. In fact, the task could easily be delegated to an algorithm. Obvious limitations arise, however, if the annotation is incorrect, or if a drawing bears a label not previously encountered, such as "Bozif" or "Zirob."

The second strategy consists in laying out all the drawings before you, so you can see them in their entirety, and grouping together those that display the most statistical similarities. You thus form two piles based on your intuition. If you are a normally functioning human being, you will likely notice that some shapes have rounded contours, while others exhibit jagged edges. You will then separate them accordingly. All that remains is to determine which pile corresponds to the Bobog you were given as an example, and which to the Zirf. To solve this final step, you again rely on statistical resemblance. The entire methodology thus consists in skillfully exploring two questions: Which shapes bear the greatest resemblance to one another? And which shapes most closely resemble the Bobog and the Zirf you were shown at the start?

This approach, as just described, does not rely on your colleague's annotations. However, it may lead to unexpected results. In examining which figures most resemble each other statistically, you might find yourself grouping together those drawn in blue ink and those in yellow ink. You might also end up associating large shapes with small ones, or separating those drawn with solid lines from those traced with dotted lines. In truth, there

is no single correct answer. In other words, statistical association is not of the same order as the rational approach previously described. Instead of relying on rational judgements defined by unambiguous linguistic rules, the synthetic judgement draws upon considerations grounded in experience. Neural networks operate in this way and, lacking intensive training, they are prone to error.

The third and final strategy consists in examining the drawings one by one and asking, in each case, whether it resembles the generic idea of a "Bobog" or that of a "Zirf." This generic idea is formed by subtracting from the initial model all that does not seem essential. You disregard size, colour, and line thickness, focusing only on what appears to constitute the most significant contrast: Bobogs are made up of curves that bulge outward (convex curves), whereas Zirfs are built from curves that bulge inward (concave curves). This third strategy thus invites you to construct in your mind a generalized archetype of the Bobog and the Zirf by abstracting their salient properties. This process is based on judgements that are indeed true, but not fully justified. There is, to be sure, an effort to move beyond mere observations of experimental resemblance, yet without being able to rely on a rational definition of what a Bobog or a Zirf actually is.

I can already hear the reader voicing a crucial objection: is the third strategy truly different from the second? This capacity for abstraction that I have described—and distinguished from the operations of neural networks—might it not also be the product of statistical associations? If we trained a machine to count the number of sides, just as a child learns to recognize geometric shapes by playing with wooden blocks, would we not end up with something essentially similar? Is there truly, within geometric abstraction, an element that transcends the play of statistical experience?

This objection, which is of central importance, will receive multiple responses over the course of this book (see, in particular, chapter 8.1). One may especially highlight the difference in learning. To create an algorithm capable of identifying complex objects or producing text like ChatGPT-4, one must train it on more data than a human being will encounter in an entire lifetime. The results, though impressive, do not always rival human judgement. A famous meme, which circulated widely on social media in 2022, illustrated how an algorithm might confuse a chocolate muffin with the tearful face of a chihuahua… or a mere apple core with the watchful eyes of a barn owl (see figure 2.3). It seems that neural networks—though they surpass human performance in a wide range of tasks—do not necessarily operate on the same fundamental principles as human cognition. The latter, despite its fallibility in logical analysis, is capable of grasping concepts based on a very small number of examples.

Figure 2.3: Neural networks, lacking intensive training, make identification errors more frequently than an adult human when faced with this kind of image.

2.1.1 The Synthetic Judgement *a priori*

The categories manipulated by the human mind are *intuitive concepts*. The word *concept* suggests that something in them belongs to the realm of reason, while the word *intuitive* reminds us that these concepts are also grasped through the inspection of our senses. The kind of judgement that allows us to isolate intuitive concepts, although quite ordinary in daily life, contains something utterly extraordinary from a philosophical perspective. It is a type of judgement that does not pertain to objects of concrete experience (observable geometric forms), but to objects of possible experience (the generic idea of a Zirf or a Bobog).

What do I mean by that? That this kind of judgement rests on a significant *a priori*: it assumes that the Bobogs and Zirfs we have before our eyes are merely possible instances of a more general structure. In other words, we suppose that what we determine to be true or false about these particular cases teaches us something about all analogous objects. Put differently again, we are incapable of observing a geometric shape as seemingly random as a Bobog without mentally referring to all sorts of other possible geometric forms. We compare the Bobog to circles, squares, triangles stored in memory, and we seek out similarities, contrasts, points of rupture.

The phenomenon I'm describing becomes especially visible in everyday social life. When you try to grasp the character of a person you've just met, your mind leaps into all kinds of conjecture. To make this more tangible, imagine you've matched with a young man or woman on your favourite

dating app and have agreed to a first meeting. From that moment on, your mind is seized by the question: who is this person? Do they match the image you formed from reading their profile? Upon seeing them for the first time, hearing them speak, discovering their internal logic, you begin comparing them to people you already know. In fact, it may be impossible to discern the personality of a stranger without assimilating them to all kinds of people they are not, but who nonetheless resemble them. You compare them to real individuals you've met, but also to fictional characters, to archetypes. These are so deeply embedded in our cultural imagination that language itself overflows with antonomasia: a miser is called a Harpagon, a seducer a Don Juan or a Casanova, a genius a Da Vinci or an Einstein, and so on. A languid woman, for those familiar with Russian literature, is likened to Anna Karenina. She is the Russian equivalent of the French Emma Bovary, with a distinctly Slavic hint of the "cursed poetess."

Thus, we might say that human judgement, when it manipulates intuitive concepts, is indeed *synthetic*, but also *a priori*. It is *synthetic* because it expands our knowledge by making new associations (it sheds light on someone's personality, on the properties of certain geometric forms, etc.). Yet it is also *a priori* because it depends on a set of presuppositions: we assume in advance that the person we are observing, or the figure we are studying, partakes of an archetype. This archetype depends on how we structure information and represent the world. It is therefore shaped both by our cultural concepts and our neurological predispositions.

Synthetic *a priori* judgement sometimes teaches us more about our own perspectival filters and ways of connecting information than it does about the objects it aims to grasp. That is why, when person A makes a claim about person B (that B is miserly, seductive, manipulative...), we learn just as much about the mental universe of person A as we do about B's behavior. To put it a bit poetically, the synthetic *a priori* judgement is the space in which associations are formed between known and unknown things, between real objects and possible ones, between experiential truths and *a priori* forms. It makes visible the links that bind our concepts together. It reveals the deeply interdependent structure of the ideas we handle.

The simplest and most elegant historical example of a synthetic *a priori* judgement can be found in Thales's demonstration that "the diameter of a circle divides that circle into two equal parts." At first glance, this proposition appears utterly unremarkable. Trying to prove it seems as futile as pushing open an already open door. And yet, what is truly remarkable in this proof is not the conclusion, but the method. For the first time in the history of mathematics, we witness a deliberate effort to establish a valid demonstration for all *possible* circles and diameters. Rather than simply verifying the property by drawing circles with a compass, the goal is to universalize the reasoning, so that the objects being manipulated are no longer this circle or that circle, but the circle in general — the *archetypal* circle.

The proof Thales suggested uses what is known as a proof by contradiction. The idea is to assume that what one seeks to prove is false, and then show that this leads to a contradiction. According to the principle that a proposition must be either true or false — the law of the excluded middle — we are then forced to accept that, if it cannot be false, it must be true. A clear explanation of Thales's proof can be found in the Intellectual Mathematics podcast[1]

> Suppose the diameter does not divide the circle into two equal halves. Very well, we now have a line that goes through the centre of a circle, and this circle is divided into two parts. And we suppose that these two parts are not identical. Take one of the parts and flip it over onto the other, as if you were folding an omelet or a crepe. We assumed the two parts were not equal, so by superimposing one onto the other, they do not match up perfectly. There must then be a spot where one part sticks out beyond the other.
>
> Now draw a radius in that direction, from the centre of the circle to the point on the perimeter where the two halves don't align. This means one radius is longer than the other. But that implies the figure wasn't a circle to begin with. A circle is a figure in which all points on the perimeter are at the same distance from the centre in every direction. That is what it means to be a circle.

In this proof, the human mind does not simply attempt to establish statistical relationships between figures drawn on paper. Instead of looking at circles and diameters with the eyes, it appeals to generic images. The generic image of a circle folded onto itself does not refer to a *particular* experience, but to a *possible* one. What I mean is that the thought experiment through which you carry out this operation is valid for all circles, whether real or imaginable. You know, then, that even if the circle you mentally picture has a specific size, that size does not matter. In other words, you possess a universal notion of the circle — one that applies to all conceivable circles.

For this reason, we can say that the conclusions you reach are *a priori* in a double sense: they are both universal and arbitrary. Universal, first of all, because they hold true in every possible experience. Regardless of the diameter, the colour, the thickness of the lines, or the position on the page, you will invariably reach the same result.

These conclusions are also arbitrary, because we cannot be certain that a perfect circle exists in reality. If you draw a circle on paper, you'll notice it

[1] Translated from the English episode *First proofs: Thales and the beginnings of geometry*, May 16, 2020, https://intellectualmathematics.com/blog/first-proofs-thales-and-the-beginnings-of-geometry/

is never perfectly circular — there are bumps, slips, inaccuracies — so that Thales's demonstration fails in practical execution. It is thus somewhat arbitrary to think that what you've drawn is a circle, and not just a cloud of points. One might argue that other cultures, vastly different from our own, would relate their observations to other generic ideas, to other abstract shapes than the circle we know. Their geometry would not be the Euclidean geometry familiar to us, but a set of perceptual filters as strange as they are unimaginable. I shall return to this idea later (see notably chapter 4.3).

To summarize what has just been said about synthetic *a priori* judgements, the rest of this thesis will attempt to put the following idea to the test (with mixed results):

Thèse 2. *Human beings possess a faculty of judgement that is imitable neither through pure experience (statistical associations) nor through pure logic (the manipulation of formal rules), but which consists in a form of abstraction that allows objects of sensible experience to be subsumed under a rational category. In Kantian theory (which I shall later critique), this judgement is called the synthetic a priori judgement. It refers neither to real objects nor to abstract definitions, but to possible objects.*

2.1.2 The Limits of Platonism

Plato's thought, as we have seen, is deeply influenced by the method of geometrical demonstration. It celebrates the *a priori* demonstrations of Thales and Euclid. However, it does not distinguish a third level between purely synthetic judgements and purely analytical ones. If the father of the Academy of Athens does not give a specific name to synthetic *a priori* judgements, it is because he believes that the reality of the world is populated by essences. For him, only the Idea of the triangle is true, not the ersatz figures traced in the sand. Though represented through the senses, the Ideas do not fall under their sway. They are accessible solely through the mediation of the intellect. This is why, in Platonism, all things in the world—whether the shapes in the sand, the sound of high tide, the howling wind, the sea spray, the salt and the crashing surf—all ultimately do nothing but partake in the expression, within the world, of universal essences. The sensible world is the concrete, albeit distorted, manifestation of an ideal reality.

To express Plato's thought differently: there must exist, in the intelligible world (*topos noetos*), an Idea of the triangle, an Idea of the waves, of the tide, of the howling wind, of the sea spray, the salt, and the surf. These Ideas, however, are not comparable to the objects of the sensible world that fall under our eyes. Nor are they possible objects, as in Kant's theory. Plato's Ideas are pure essences, ideal forms—precise definitions. Our soul was able to contemplate these essences before being born into a body, and retains a

memory of them still. When this soul says of a vaguely traced shape in the sand, "this is a triangle," it does nothing but retrieve a primordial memory, a forgotten dream, an ancient and unalterable remembrance.

The method of learning that allows the soul to give birth to the truths it harbors is called *maieutics*. It consists in posing the right questions so that the human being, through the use of reason alone, rises to the highest level of precision in the articulation of definitions, analytical concepts, and the relations between them. Plato offers an example in the dialogue *Meno*, where Socrates, by skillfully questioning a slave, manages to guide him toward an understanding of geometrical truths.

This is why the kind of judgement that consists in saying "look, it's a triangle!" does not concern Plato in the first instance. It falls on the side of *Pistis* (belief) and holds no privileged epistemological status. This judgement is nonetheless discussed in the *Theaetetus*; I have pointed this out on several occasions. It is even this judgement that lies at the origin of the notion of "Understanding." This word, which appears in the writings of Descartes, Spinoza, Hume, and Locke, designates a faculty occupying a position intermediate between Sensibility—focused on sensible objects—and Reason, which contemplates the pure intelligible forms.

This notion of "Understanding" is of central concern to us, for it seems to be precisely one of the points where artificial intelligence still encounters notable difficulties. It is easy enough to train a machine to make logical deductions, drawing a rational conclusion from a rational premise; it is likewise possible to ask an algorithm to identify patterns within sensory data. Yet it remains difficult to generalize sensory observations in a way that allows them to be manipulated abstractly. In other words, artificial intelligence, in comparison to the human being, struggles to form categories that are operable by rational thought.

In the chapter titled *Artificial IQ* (chapter), I illustrated certain types of problems where machines still perform poorly. These problems, invented by François Chollet, resemble Raven's Progressive Matrices. They consist in generalizing certain patterns from a small number of examples (see figure 2.4). Tasks that require flexible logical structures and do not rely purely on statistical imitation of large datasets are likely to pose a challenge to current programming paradigms. These logical structures that I have in mind often involve distinguishing cause from effect, identifying unity from plurality, recognizing symmetries, and so on. One might say that these operations of thought are a kind of judgement—an act of breaking down the initial image, taken as a whole, into meaningful individual elements.

Because of its high degree of abstraction, this kind of judgement appears to occupy a position between what Plato called *Pistis* (belief) and *Dianoia* (deductive logic). It is a judgement that is indeed synthetic, in the sense that it groups together sets of pixels under a single unity, but not entirely *a posteriori* (dependent solely on experience, without the intervention of reason).

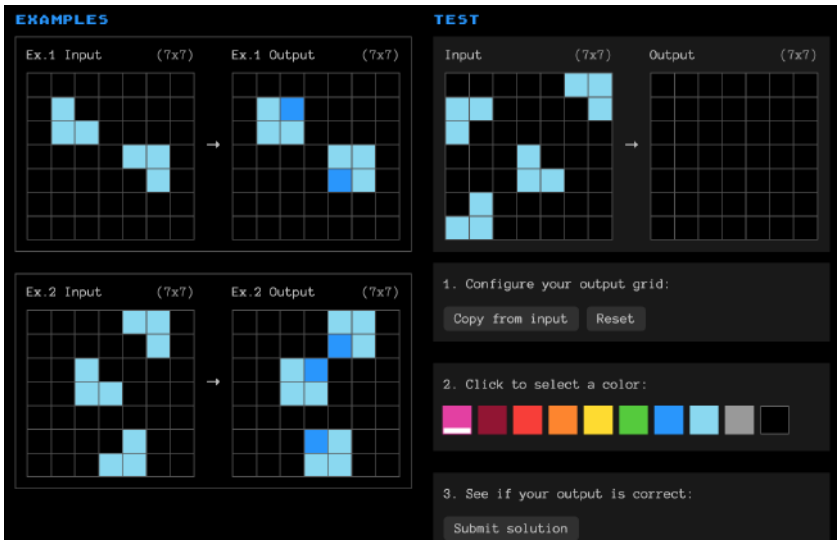


Figure 2.4: In this Raven-like matrix, the machine is given fragments of squares with a missing pixel, and shown—by means of only two examples—that it must complete the square with a blue pixel. The machine must then apply what it has just learned to new situations.

Far from relying exclusively on experience, it must partially rest on abstract laws, properties of symmetry, and spatial and temporal intuitions—all of which stand in sharp contrast to the vast statistical correlations of big data.

Since it is neither a synthetic *a posteriori* judgement (it does not depend exclusively on experience), nor a purely analytical one (it does not depend solely on definitions), it seems to hinge on another ingredient, beyond those we encountered while studying Plato's thought. In the following pages, I would like to suggest that this ingredient might be comparable to the synthetic *a priori* judgement introduced by Immanuel Kant in his monumental *Critique of Pure Reason*.

The choice to speak to you about Kant is no arbitrary decision. Not only is he the philosopher who most emphasized the relationship between human judgement and experience—shedding light on the existence of an intuitive capacity for abstraction—but he is also known for having established a close dialogue with Plato's theory. He sought to draw philosophy out of that kind of speculative rationalism, those liminal realms of ideality in which the Greek gods seemed to have birthed it.

In the chapters to come, I will focus especially on the *Critique of Pure Reason*, endeavouring to highlight several passages that refer to the *Theaetetus*. I will attempt to introduce the definition of the synthetic *a priori* judgement through a comparative analysis with the Socratic dialogues, after which I will show how these dialogues contribute to the introduction of a new human faculty: Understanding. These discussions will prepare the

ground, I hope, for a deeper inquiry into my main hypothesis: is the synthetic *a priori* judgement precisely the hiatus, the unknown, the missing link that artificial intelligence lacks in order to solve abstraction exercises similar to François Chollet's ARC tests?

It is important to note, however, that my thesis will not be inscribed purely and simply within Kant's theory. I introduce here the concept of the synthetic *a priori* judgement as a step toward constructing, later on, a theory of mind capable of accounting for why we are able to form a mental image of the "circle in general" or the "diameter in general," and how we can manipulate these concepts to establish relationships.

2.2 Kant vs. Plato: The Trial of Ideas

Kant takes up philosophy exactly where Plato had left it, but rather than blindly extending Socratic dialectic to metaphysical questions, instead of addressing the question of what makes a (universal) science of the sensible and its transformations possible (as in the *Theaetetus*), he wishes to first raise a methodological doubt: what is it possible to know through reason? It would, in his view, be negligent to conquer the intelligible world without first understanding what grants us access to it. For before using a tool, even the most powerful at our disposal, one must first analyze it. This is undoubtedly why he says that:

> *If it had occurred to any ancient thinker to pose this question, it would have erected a powerful barrier to all the systems of pure reason that have arisen up until now, and it would have spared many futile attempts, which were made blindly without properly understanding what was at stake*[2].

What exactly are these attempts that Kant condemns? What are the criticisms he directs at Western philosophy and metaphysics in particular? To get a clearer idea, perhaps it is helpful to recall that Plato introduced Reason in such a way as to enable a new type of judgement, one that is both universal and synthetic. In doing so, he led philosophy into the realm of pure thought—a thought that no longer deals with sensible impressions, but with universal concepts such as numbers. This allowed him to envision a knowledge of the intelligible that, like mathematics, expresses itself through sensible images, but without depending on them. A knowledge that transcends the sensible and gradually rises to the anhypothetical.

Plato's conception, alas, comes at the cost of a certain dogmatism: the existence of the Forms and the transcendence of the Idea of the Good (about

[2] KANT [99], *Critique of Pure Reason*, trans. BARNI, Introduction V, note (a) on page 58.

which one may wonder what it precisely looks like). While we must be indebted to Plato for insisting on the existence of a priori knowledge, we must also criticise him for failing to provide a genuine *thinking* of pure thought. According to what criterion can we determine whether a piece of knowledge reflects an Idea, or if it is merely a misguided opinion? What are the properties of pure thoughts, and how do they relate to one another? In other words, what are the rules of reasoning? Plato indeed inquired into the existence of false opinion, into what distinguishes it from true opinion, but always in a rather aporetic manner. He seemed eager to build a rational edifice capable of elevating human knowledge to the peak of its power, without first verifying its foundations.

In his introduction to the *Critique of Pure Reason*, Kant intensifies his irony towards the theory of Ideas: "The light dove, which, in its free flight, cuts through the air whose resistance it feels, might imagine that it would fly much better in the void. Thus, Plato, leaving the sensible world, which confines intelligence within such narrow limits, ventured, on the wings of ideas, into the empty spaces of pure understanding[3]." Although he does not elaborate on the meaning of these few lines, their significance is revealed through numerous other references, more or less direct, to the theory of Ideas[4].

Indeed, it is likely that the "light dove" directly refers to the pigeon loft analogy in the *Theaetetus*[5]. We have already discussed it in the section titled *True Judgement* (section **??**), but without delving into all the details. Since it is of considerable importance, I will take care to discuss its limitations here in greater detail, while adopting a Kantian perspective throughout. This will provide an opportunity to highlight the fundamental differences between the two authors. It will also introduce the project of the *Critique of Pure Reason*, in terms of its innovative approach to the perspective on synthetic propositions.

2.2.1 The Aviary Metaphor

In the *Theaetetus*, a theory of knowledge is presented through the metaphor of the dovecote. This allegory is intended to account for the possibility of

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

[4] See especially the first section of the transcendental dialectic: "Plato used the word idea in such a way that it is clear he meant something that not only does not derive from the senses, but even surpasses the concepts of understanding addressed by Aristotle, since nothing could be found in experience that corresponds to them. Ideas, for him, are the types of things themselves, not merely keys to possible experiences like the categories. In his opinion, they derive from supreme reason, from which they passed into human reason, but the latter is no longer in its original state, and it is only with difficulty that it can recall, today quite obscured, its ancient ideas, through reminiscence (which is called Philosophy)" (Kant [99], *op. cit.*, p. 311, also pp. 312-315).

[5] See *Theaetetus* 197b-200e.

making errors of judgement within pure thought. To recall, it consists in likening knowledge and ignorance to wild birds[6]. In this context, to learn means to capture the appropriate bird and place it in the labelled cage of a grand dovecote. To remember and apply knowledge that has been acquired is thus to search within one's cage for the bird that holds that knowledge.

According to this conception, error occurs when one takes a bird for another, or mistakes ignorance for knowledge. This theory therefore presupposes the existence of two types of reasoning: the first are judgements relating to sensory impressions, the second are judgements of pure thought. According to Socrates—and through him, Plato, who holds the pen—a mental operation such as "5 + 7 = 12" is an act of pure thought[7], which, as such, refers to no sensory intuition whatsoever.

It may be useful to ask where such a theory of pure thoughts leads us. Since some doves, instead of carrying genuine knowledge, bear ignorance, there must exist somewhere a knowledge concerning these doves, allowing us to determine whether what they carry is true or false. This means nothing other than this: for each dove, it must be possible to know whether what it says is valid by referring to the information held by another dove. The relevance of what this second dove has said must, in turn, be verified by a third dove, and so on. In this respect, we are quickly drawn into an infinite regress.

In more contemporary terms, the problem raised by Socrates is thus that of what makes the reflexivity of knowledge possible. Either one admits that every thought, by the very fact that it is a thought of something, is necessarily true, somewhat in the manner of Parmenides—but in that case, one struggles to explain the origin of error, and risks, along with Protagoras, sinking into the meanders of relativism. Or, conversely, one admits the possibility of false thoughts, but then there must also be a thought of false thoughts, and then a thought of the thought of false thoughts, and so on, such that the final verification of validity can never be provided. To seek truth under these conditions is to pursue the ultimate term of an endlessly chained sequence of conditions. It is to seek to ascend to the unconditional or, in Plato's terms, to the anhypothetical.

[6] The theory of the dovecote is developed throughout the dialogue. In a first version (*Theaetetus* 197b–197e), each dove carries a piece of knowledge, and error thus consists in confusing one bird for another. Socrates then points out a troubling issue: the error always stems from the fact that the subject knows something, but confuses what he knows, and not from pure ignorance.

To correct the model, Theaetetus proposes a second version (*Theaetetus* 199d–200a). In this one, the birds may carry either knowledge or ignorance. Error can then result either from the confusion between several knowledges or from taking ignorance for knowledge. It is this second version that I examine here, even though, for the sake of simplicity, I had presented the first version in section **??**.

[7] *Theaetetus*, 196a.

This objection is raised by Socrates himself[8], yet it does not lead him to revoke pure thought, nor to limit its use. On the contrary, the dialogue continues with a new definition of science: that of "true opinion accompanied by reason"[9]. Reason is thus immediately posited as the touchstone of pure thought, as the reliable criterion meant to safeguard against both error and infinite regress. Now, Kant's perspective consists precisely in extracting true opinion from reason—and therein lies the inaugural nature of his gesture.

At this point, a scrupulous reader might object that the dovecote objection does not concern Plato's own theory, but likely some theory defended by his contemporaries. The dialogue, one might argue, was not meant to establish this representation of pure thought, but rather to point out its dead ends. This observation is entirely legitimate. That is why I would like, to forestall any hesitation on this point, to show that similar difficulties arise in the theory of Ideas.

To see this, it suffices to ask the following question: is there, somewhere on the edge of the intelligible world, an Idea of the False or an Idea of Error? If one answers in the negative, this implies that the only way for a pure thought not to be true is for it to participate in no Idea. But a pure thought that does not participate in any Idea, in Plato's language, is not a thought at all. Error, then, cannot be within pure thought, but only within the sensible. This position makes it difficult to explain how mistakes can be committed in abstract reasoning, such as mathematical reasoning.

If one answers in the affirmative, and admits an Idea of Error, this leads to no more satisfying a conclusion. For if there truly exists an Idea of Error, then everything that participates in that Idea is at once true (since it relates to an Idea) and false (since that Idea is the Idea of Error)[10]. The notions of true and false, if they exist, must therefore be said in several senses. This problem evokes in many respects certain results of contemporary logic[11], and may well be the prelude to the many paradoxes of self-reference.

What, in Kant's eyes, lies at the root of such paradoxes? As long as knowledge concerned sensible objects, one had at hand a criterion of verification, an experience against which to measure the adequacy in question; but as soon as one leaps toward pure thought, reason has no foundation other than itself, and adequacy can no longer be conceived without immediately becoming ensnared in contradiction. This is, fundamentally, Kant's most profound critique—addressed to Plato as much as to all the metaphys-

[8] *Theaetetus*, 199b–200c.

[9] *Theaetetus*, 201d.

[10] Analogous objections are raised by Aristotle in the ΠΕΡΙ ΙΔΕΩΝ (or what has been preserved of it thanks to commentators such as Alexander of Aphrodisias), especially the idea that there would have to exist an Idea of non-Being. For a more thorough discussion of these arguments, see for example Mansion, *La Critique de la Théorie des Idées dans le ΠΕΡΙ ΙΔΕΩΝ*.

[11] I am thinking in particular of Tarski's theorems. The chapter entitled *Crisis in the Foundations of Mathematics* will explore this question in more detail.

ical projects that followed in his wake. Yet far from merely rejecting the ontological project like an impartial judge, he undertakes to rehabilitate it. How? By limiting the use of the understanding to the objects of possible experience.

Some might be inclined to think here that the rejection of pure thought entails the abandonment of mathematics. This would be to misunderstand the originality of Kantian thought. Unlike Socrates, who held the proposition "5 + 7 = 12" to be a judgement of pure thought[12], Kant sees in it a judgement of the understanding grounded in intuition[13]:

> *One is no doubt tempted at first to believe that the proposition 7 + 5 = 12 is a purely analytic one, resulting, according to the principle of contradiction, from the concept of the sum of seven and five. But on closer inspection, one finds that the concept of the sum of 7 and 5 contains nothing more than the uniting of two numbers into one, and that it in no way lets us know what this single number is that contains both. [...] I must go beyond these concepts by having recourse to the intuition that corresponds to one of the two, for example to the five fingers of the hand, or (as Segner teaches in his arithmetic) to five dots, and in this way gradually add the five given units in intuition to the concept of seven*[14].

What exactly is the nature of the intuition to which Kant refers here? It is a pure intuition of space and time, inherent to the very structure of subjectivity, which exists independently of any sensory experience. Kant's position inaugurates mathematical intuitionism and occupies a middle ground between Plato's mathematical realism and empiricism[15]. In his view, mathematical truths do not refer to intelligible objects existing independently

[12] In the *Theaetetus* (196a), Socrates raises the question of whether a miscalculation such as "5 + 7 = 11" stems from an error in pure thought or from a mismatch with sensible experience. Several pages later, he declares: "We must therefore show that false opinion is something entirely different from a disagreement between thought and sensation. For if that were all it was, we would never be mistaken in our pure thoughts." (*Theaetetus*, 196c). This indicates that arithmetic propositions, for Plato, are nothing but judgements of pure thought.

[13] The fact that Kant exemplifies the synthetic *a priori* judgement with the proposition "5 + 7 = 12," which is found as such in the *Theaetetus* (*Theaetetus*, 196a), once again shows the close dialogue he maintains with Plato throughout his critique.

[14] Kant, *op. cit.*, Introduction V, pp. 59–60.

[15] Kant, in fact, often defends himself against both idealism and empiricism. See, in particular, the section "Refutation of Idealism" in the Transcendental Analytic (Kant, *Critique of Pure Reason*, trans. Tissot, pp. 238–248), or again the "Refutation of Mendelssohn's Argument for the Permanence of the Soul," which concerns psychological idealism (Transcendental Dialectic, pp. 340–348). For empiricism, see also the discussion of Locke and Hume in the "Transition to the Transcendental Deduction of the Categories," (Transcendental Analytic, pp. 134–136).

of the subject in a realm of idealities. Nor are they the mere product of mental construction; nor, again, are they empirical knowledge drawn from experience of the sensible world. They are the *a priori* structure of all possible experience—or, in other words, they are the very forms of all reasoning concerning sensible objects.

2.2.2 Knowledge *a priori*

This refoundation of metaphysics is further motivated by another aspect: making *a priori* knowledge possible. Let me explain. In Plato, the Ideas are the supreme objects of thought. These Ideas exist independently of the human mind. They must therefore already be given in order to be contemplated. Under such conditions, it becomes difficult to maintain the possibility of strictly *a priori* knowledge, such as that of mathematics. This is why, according to Kant, the genuine method in speculative philosophy consists in positing that objects conform to human knowledge, and not the other way around.

This reversal of perspective is nothing short of a Copernican revolution: "If intuition had to conform to the nature of objects, I do not see how anything could be known *a priori*; but if, on the contrary, the object (as an object of the senses) conforms to the nature of our intuitive faculty, then I can very well understand how such knowledge is possible[16]." Everything that is the object of *a priori* knowledge is always already within us, prior to any sensible or intellectual experience. With Kant, the Platonic *noetic* synthetic judgement is replaced by the *a priori* synthetic judgement.

A major consequence of this method is that it is no longer possible to know things other than as phenomena. Since it is the objects that must conform to the nature of our concepts, and not the reverse, we only ever know them through the modality of our perception. As a result, it is impossible to extend knowledge beyond the domain of possible experience. Yet within humanity there exists a natural disposition, in the ongoing development of knowledge, to pose questions that transcend all representation rooted in sensible intuition. These are the questions concerning God, the world, and the soul. Is the world finite or infinite? Does it have a beginning in time? Was it created by a Supreme Being? Is the soul a simple and incorruptible substance? These are the interrogations that have nourished every system of belief, whether metaphysical or cosmogonical. Since such ideas refer to no possible experience, they always lead to paradoxical reasoning and must therefore be set aside. It is the task of *transcendental dialectic* to carefully examine each of the logical conclusions to which these considerations lead, so as to reveal their inextricably aporetic nature.

[16] KANT [99], *Critique of Pure Reason*, trans. by BARNI, Preface to the Second Edition, p. 24.

Upon close examination, the first result of the critique undertaken by the philosopher is essentially a *negative* one: it is to show that the use of pure reason cannot lead to genuinely scientific knowledge. Kant therefore proscribes Plato's *noēsis*, understood as the search for the anhypothetical. This restriction, however, far from toppling dialectic from the pedestal on which Platonic philosophy had placed it, contributes on the contrary to reinforcing its position. It must, of course, descend a few steps; it must give up its claim to embrace the totality of knowledge in a single gaze — and yet the practice of dialectic as such is preserved: "there has always been, and always will be, a metaphysics in the world, but alongside it there will always arise a dialectic of pure reason, for it is natural to it. The first and most important task of philosophy is therefore to strip this dialectic of all pernicious influence once and for all, by destroying the source of its errors[17]."

The Reason inaugurated by Plato, though now compelled to greater modesty, also survives the critique. First of all, a certain use of human intelligence — namely, that of knowing the objects of possible knowledge *a priori* — is preserved and even consolidated. This use is based, on the one hand, on sensible intuitions, which provide the subject with the matter of knowledge, and, on the other, on a form of discursive reasoning that allows one to categorize the sensible data. It bridges the gap between sensibility and the intelligible, without getting lost in pure thought. It thus remains quite close to the discursive reason (*dianoia*) of which Plato spoke, and perhaps it may even be seen as its continuation? Kant, and before him a long metaphysical tradition, referred to this faculty as the "understanding". Its domain is that of *a priori* synthetic judgements, which encompass all of pure mathematics and physics.

Moreover, the restrictions imposed on pure reason do not extend to practical reason. On the contrary, Kant is intent on preserving the moral order. Although the questions of freedom, the existence of God, and the immortality of the soul — when treated speculatively — have evaporated under the fire of critique, the question of the moral law, for its part, has survived the trial by fire. But one may wonder how such a miracle is possible. It may indeed seem surprising to ground a universal *a priori* law of morality when human freedom cannot be securely sanctioned by Reason. And yet from the embers of critique remain the ashes of belief. Freedom is indeed aporetic — but nothing prevents man from believing in it. Let us clarify this perspective. Since reason can only be exercised on phenomena, and not on things in themselves, there is nothing contradictory in thinking that noumena may be free, while also maintaining that the soul, as it appears to us, can appear only under the mode of natural causality. This point of view allows us to reconcile the antinomy in which we find ourselves: on the one hand, the subjective certainty of being an autonomous subject, and

[17] KANT, *op. cit.*, Preface to the Second Edition, p. 36.

on the other, the knowledge of being no more than an organism subject to immutable physico-chemical laws.

This leads Kant to affirm that the best proofs one can offer man to guide him on the path of faith and virtue are not the abstruse demonstrations of the scholastics, but the compelling testimonies of human experience[18]. It is vain, and illusory, to try to prove metaphysical truths — upon which the moral order supposedly depends — by apodictic means. One need only speak to the heart of man. Should he doubt the existence of God? Point out to him the harmony and unfading beauty of nature. Should he doubt the immortality of the soul? Wait until he experiences that nothing temporal can ever truly satisfy him. Should he doubt the freedom of his will? Confront him clearly with his duties as opposed to his inclinations, and awaken in him the consciousness of choice. What is threatened by the Kantian critique is not the moral or religious principles themselves, but the monopoly of schools of thought that uphold these principles dogmatically.

Note that the process by which *noēsis* is disqualified in favor of *dianoia* is of crucial importance for understanding what the concept of Reason has come to mean in the Western world — especially at the height of the 20$^{\text{th}}$ century. *Dianoia* operates by selecting a set of hypotheses, by decomposing concepts into smaller elements, by calculating results in a mechanical way. It belongs to an economy of thought that is that of discursive logic, of analytical engineering. It is therefore no surprise that Plato's Reason, which united both a discursive and a contemplative dimension, has become what Weber called instrumental rationality[19].

It is not by chance, then, that critiques of this instrumental rationality — as found in postmodernism or in phenomenology — often involve a revalorization of the non-discursive dimension of reason. This can be achieved, and has often been, in two ways: either by rehabilitating *noēsis*, that is, by returning to the intuitive or contemplative functions of reason; or by rehabilitating *practical reason*, through the recognition of norms, systems of values, and meaning in the orientation and organization of rationality[20]. I will give two examples to support this idea, both drawn from the Frankfurt School[21].

The first example, which in my view illustrates a return to a form of *noēsis* — though a *noēsis* freed of its metaphysical connotations — may be found in the concept of *mimesis* developed by Adorno and Horkheimer[22].

[18] *Ibidem*, p. 37.

[19] I offer this idea as a hypothesis. A more detailed analysis would take us beyond the scope of the present work.

[20] This again is a personal hypothesis.

[21] The Frankfurt School is a movement of social philosophy influenced by Marxism and the Enlightenment ideal, which undertakes to critique capitalism in order to promote the emancipation of individuals. Key figures include Max Horkheimer (1895–1973), Theodor W. Adorno (1903–1969), and, more recently, Axel Honneth (b. 1949).

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

According to these authors, contemporary rationality has developed within a logic of identity: first, by departing from poetic discourse, that is, by severing the bond between sign and thing; and second, by subsuming the diversity of the Other under universal concepts. This has led to a double domination: that of external nature, objectified and reduced to a set of exploitable resources; and that of man's inner nature, whose creativity is repressed. In response to this situation, they call for a return to a form of mimesis — not to reappropriate nature and remake it in our image, but rather to make space within the self for its otherness. The task of subjectivity is no longer to identify the non-identical with itself, but to identify it as it is.

The second example is that of the lifeworld, as conceived by Habermas. If the contemporary world is dominated by a functional-instrumental rationalization, it is due to what he calls a selective rationalization. Yet another form of rationality exists: communicative reason. It is defined by an exit from strategic logic. Rather than seeking to influence the other, to use them as a means to an end, one must confront one's own point of view with theirs, following the law of the better argument, until consensus is found. Such a rationality, insofar as it is exercised through language within a community, and insofar as it is only possible through processes of interaction, is also by essence a form of action. Habermas calls it communicative action. That said, it is not a matter of bringing all spheres of human activity under the imperative of consensus — for if that were the case, society as a whole would be compelled to justify at every moment the reasons for its existence; it would lose efficiency, cohesion, and integration. On the contrary, instrumental and functional reason must continue to govern society as a system — but it must leave the world of norms, meanings, and values to communicative reason. This world is the lifeworld.

Now that Kant's critique of Plato is, we hope, clearer in the reader's mind, it is time to detail the contents of the *Critique of Pure Reason*, so as to better understand the status of the *a priori* synthetic judgement, and, in particular, the nature of the synthesis at stake here.

2.2.3 The Copernican Revolution

In Kant's view, the history of the sciences has endowed mathematics and physics with methods appropriate to their objects, to which nothing can be added or subtracted[23]. As for the mathematical method, it was discovered at the moment when Euclidean geometry was born, at the moment when the conviction emerged that mathematical propositions could be the subject of apodictic demonstrations. What do we mean by this?

The great idea of the Greeks, what allowed them to emancipate them-

[23] KANT, *op. cit.*, Preface to the Second Edition, pp. 19-22.

selves from Egyptian or Sumerian mathematics, was their ability to think about the properties of geometric figures without feeling the need to represent them in space[24]. What science says about the triangle is what it says about the essence of the triangle, that is, about every triangle, whatever it may be. By doing so, its judgements are perfectly universal and *a priori*. Euclidean geometry achieved such perfection that it is no longer possible to advance it further. At most, one can present its results more clearly, or include it within a more general theory as a particular case, as is the case with the position currently occupied by Euclidean geometry within non-Euclidean geometries.

In the case of physics, the method took longer to emerge into the light of day. And for good reason, for its object is deeply intertwined with sensory experience, so much so that it was necessary to determine how the principles of reason should accommodate observations of nature, and vice versa. It is to Bacon, to whom Kant also dedicates an epithet[25], that Kant attributes the paternity of the correct method. This method had already been glimpsed, as if through a play of reflections, in the experiments of his predecessors: Galileo's inclined planes, Torricelli's barometers, and the controversial theories of Sthal. Reason gradually learns to recognize and direct its intervention in experience. What it perceives in an experimental setup is nothing other than what it itself has put into it according to its own plans. Far from simply receiving nature's words as a schoolchild drinks in the teachings of their master, reason must behave as a judge, endowed with the right to interrogate and, if necessary, to constrain its witnesses. Through experiments, nature is called upon to account for what it is. There is a close relationship between the principles of theory and the design of the experimental setup.

In contrast to the cases of physics and mathematics, Kant says, metaphysics has never made real methodological progress, for it "has not yet been sufficiently favored by fate to enter the sure path of science[26]". The reason for this gap, according to Kant, lies in two causes. First, *a priori* knowledge can attribute nothing to objects other than what it has itself placed there,

[24] "The first person who demonstrated the isosceles triangle (whether named Thales or something else) was struck by a great light; for he discovered that he should not focus on what he saw in the figure, or even on the simple concept he had of it, but that he needed only to bring forth what he himself had introduced into it through thought and constructed *a priori*, and that, to know something *a priori* with certainty, he should only attribute to this thing what necessarily derived from what he had placed into it, according to the concept he had formed of it." (KANT, *ibidem*, p.20).

[25] "Physics arrived much more slowly at the great path of science; for it is only about a century and a half ago that a great mind, Bacon of Verulam, partly provoked and partly, since the trace had already been laid, stimulated this discovery, which can only be explained by a sudden revolution of thought. I here consider physics only inasmuch as it is based on empirical principles." (KANT, *op. cit.*, Preface to the Second Edition, p. 21).

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

so that inquiries must be directed toward reason as a faculty, and not toward its speculative objects. On the other hand, reason forms such a unity that all its principles are interconnected. Each must be analysed under all its relations to verify the coherence of the whole.

Once these two obstacles are overcome, nothing would prevent metaphysics from becoming a science. Unlike mathematics and physics, whose objects lie outside of knowledge itself, metaphysics can embrace all that there is to know in a single sweep. As soon as it has been guided onto the right path, with the correct method, there is no reason to believe it cannot, like Aristotle's logic before it, become a complete and finished body of knowledge. It can, in Kant's view, be passed down as it is to future generations, with the only necessary changes being those not to its content, but to a clearer exposition of that content.

2.3 Sensitivity: The Man Who Feels

The great Kantian idea, the starting point of his entire inquiry, consists in seeing the results of modern science, of which Newtonian physics and Euclidean geometry represent the most eminent archetypes, as a set of synthetic *a priori* propositions[27]. Of course, this point can be contested, and it has been in the recent history of physical and mathematical sciences. Non-Euclidean geometries, in particular, have explicitly raised the question of whether truths such as "The straight line is the shortest distance between two points" are, as Kant claims, synthetic *a priori* judgements[28].

That being said, to accept, as Kant does, that scientific knowledge develops neither solely through analytic judgements nor solely through *a posteriori* judgements, is to pose the question of how the production of scientifically valid judgements can be carried out *a priori*. In particular, Kant asks what must be the element X that relates the two terms involved in a synthetic judgement with universal intent, if it is not experience? This leads him to search for the conditions of possibility of scientific experience itself.

Let us illustrate this approach with an example. The following principle: "Every effect is the product of a cause" is a synthetic *a priori* judgement used by science, whose truth does not derive from experience. The empirical approach, in fact, would ultimately lead only to the Humean result that "In all the experiments conducted so far, effects are the products of a cause", without ever asserting the universality of the principle. This is why, according to Kant, such a principle must be derivable, not from experience, but from the very structure of knowledge. It is, therefore, for Kant, a matter of unveiling the *a priori* structure of each of the faculties of knowledge, namely Sensi-

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

[28] This will be the subject of the next chapter: *The Crisis of the Foundations of Mathematics*. See in particular *Objections to Kant*, section 4.4.

bility, Understanding, and Reason. This approach, which, starting from the existence of certain types of *a priori* judgements, seeks to determine their conditions of possibility, is called *transcendental*.

According to Kant, sensibility is structured by two intuitions: the pure intuition of space and the pure intuition of time, which correspond respectively to external sense (the perception of the world) and internal sense (the perception of oneself). These intuitions are what make experience possible, in the precise sense that no phenomenon can be conceived outside of time and space. "It is not sensation", says Kant, "that gives me the representation of space, since, without this very representation, I could not represent things in this way. It must therefore already exist within me[29]." Or again: "It is impossible to represent that there is no space, though one can well represent that there are no objects[30]." The same holds for time. "One cannot eliminate time itself with respect to phenomena in general, though one can well subtract them from time by thought. Time is therefore given *a priori*. Without it, all reality of phenomena is impossible[31]." While the matter of sensibility is indeed given by sensation in an *a posteriori* manner, its form, however, is present in the mind in an *a priori* manner.

This leads to two remarkable consequences. First, things are not accessible to us absolutely, as they are in themselves, but only through the prism of these intuitions, that is, through our own subjectivity. In other words, things are knowable only as phenomena: the noumenal world is beyond the reach of sensible knowledge. Secondly, time and space do not have absolute existence. To the exact extent that time and space are *a priori* conditions of sensibility, they are nothing outside of the representations of the subject. They are not realities in and of themselves, any more than they are inherent properties of the objects of experience. These intuitions certainly have objective value, in the sense that they apply to all possible objects of experience, but they do not have absolute value. To put it in Kantian terms, they are both empirical realities and transcendental idealities:

> *Our examination of space thus shows its reality (i.e., its objective value) from the point of view of the perception of things as external objects; but it also reveals its ideality from the point of view of reason considering things in themselves, that is, abstracted from the constitution of our sensibility*[32].

[29] KANT [99], *op. cit.*, Transcendental Aesthetic, Of Space, §2, p. 77.
[30] *Ibidem*, p. 78.
[31] *Ibidem*, Transcendental Aesthetic, Of Time, §4, p. 86.
[32] KANT, Transcendental Aesthetic, Of Space, §4 Consequences Drawn from the Above, B, p. 83. Later, Kant makes a similar remark about time: «We do not admit, therefore, that time can claim absolute reality, as if, even abstracting from the form of our sensible intuition, it belonged absolutely to things as a condition or property.» (*Ibidem*, Transcendental Aesthetic, Of Time, §6 Consequences Drawn from the Above, C, p. 91)

Finally, note that space and time cannot be objects of pure thought, but can only be given through sensible representation. Just as it is impossible to think of a thing outside of space and time, it is equally impossible to think of time and space independently of the things of experience. This observation constitutes a first limit that the subject encounters when attempting to think its interiority. The subject can grasp itself, through internal sense, only as it appears to itself; that is, not as it truly is in itself, but only as a phenomenon. It has only empirical knowledge concerning itself. Consequently, a differentiation immediately arises between the perceptive subject, as it is in its interiority, and the representation it acquires of that interiority. And once the true self is not accessible to knowledge, it becomes pertinent to ask how reflective knowledge of the subject and its faculties is made possible? Is this not precisely what Kant aims to achieve in the critique of pure reason? This potential circularity in Kantian thought need not be discussed here, but I present it to make it clear that the problem of self-reference, which was identified in Plato's theory, may perhaps only be displaced with the Kantian solution.

2.4 Understanding: The Thinking Man

If sensibility is the faculty related to sensation, which makes representations possible, understanding, on the other hand, is the faculty of judgement. But what does it mean to judge? It is the act of synthesizing the various representations of sensibility in such a way as to constitute knowledge of a determinate object: as when I say, "This is a metal" or "Every metal is a body[33]." Thus, the pure intuitions of space and time are not sufficient for knowledge. It is only through the action of understanding that the diversity of sensibility can be unified under a concept. To summarize this in a pithy sentence: without sensibility, nothing would be *given*, but without understanding, nothing would be *thought*.

Kant distinguishes between general logic, whose role is to define the pure forms of thought independently of its origin or its relation to an object, and transcendental logic, which, in turn, seeks to delimit the objective value of an *a priori* concept of understanding. One could say that general logic is purely formal: it studies how concepts are linked together in accordance with the laws of understanding, without inquiring into the empirical or intuitive origin of these concepts and laws.

Transcendental logic, on the other hand, raises the question of what guarantees the validity of a judgement about an object in an *a priori* manner. Kant, however, warns us not to mix in psychological considerations, but to remain strictly within a pure theory of knowledge. In transcendental logic, the question is not what empirical conditions must be met to produce a

[33] *Ibidem*, *Analytic of Concepts*, Chapter I, section 1, p. 128.

valid judgement, such as attention, enumeration, etc., but rather what the universal structures of understanding are that make judgement possible.

General logic serves as the starting point for transcendental logic. In order to determine the *a priori* structure of understanding, Kant first examines the logical functions governing judgements. These functions are simply operations of thought: categorical, hypothetical, disjunctive judgements, etc., and resemble a first-order modal logic[34]:

1
Quantity of Judgement
General
Particular
Singular

2
Quality
Affirmative
Negative
Indefinite

3
Relation
Categorical
Hypothetical
Disjonctive

4
Modality
Problematic
Assertoric
Apodictic

Kant applies these functions to the pure intuitions of space and time, thus obtaining the categories of understanding, which are nothing more than the *a priori* concepts that preside over the unification of the diversity of sensibility. Among these principles, we find those of quantity, quality, relation, and finally, modality[35].

[34] *Ibidem*, *Transcendental Analytic*, Chapter I, section 2, §9, p. 128.
[35] *Ibidem*, *Transcendental Analytic*, Chapter I, section 3, §10, p. 138.

1
Quantity
Unity
Plurality
Totality

2
Quality
Reality
Negation
Limitation

3
Relation
Substance — Accident
Causality and Dependance
Communauty

4
Modality
Possibility — Impossibility
Existence — Non-existence
Necessity — Contingency

2.4.1 Kant vs François Chollet

To provide a practical insight into these categories of understanding, it may not be useless to draw a parallel between Kant's theory and the four abilities of general artificial intelligence identified by François Chollet in chapter . As a reminder, François Chollet sought to determine which cognitive operations were necessary for a machine to succeed in the ARC exercises, a kind of artificial IQ test defined by a set of generalization problems similar to Raven's matrices (see figure 2.5). In this context, he distinguished four key ingredients: (1) object persistence, (2) geometry/topology, (3) arithmetic, and (4) goal orientation. I will recall the definitions of these four abilities and show how they resemble or differ from Kant's categories of understanding.

Arithmetic is the ability to sort objects based on their shape, colour, appearance, or movements, using addition, subtraction, and comparison. It notably involves three operations of thought: (1) the ability to distinguish unity (a pixel is green), (2) the ability to distinguish plurality (there are several green pixels), and (3) the ability to consider totality (all pixels are green). These operations therefore consist of abstracting a particular object from a whole, comparing it to the other elements of that whole, and identifying which characteristics distinguish it. These comparisons correspond precisely, in Kantian criticism, to the categories of Quantity, which allow us to make universal judgements ("all squares are green or yellow"), particular judgements ("there are yellow squares"), and singular judgements ("this square is yellow"). It is therefore possible to think that arithmetic, as a mathematical skill that can be trained, relies on a more fundamental, innate structure of human understanding. The use of the word "innate," however, does not imply that this structure exists as such from birth, but rather that it is a faculty that unfolds according to a predefined develop-

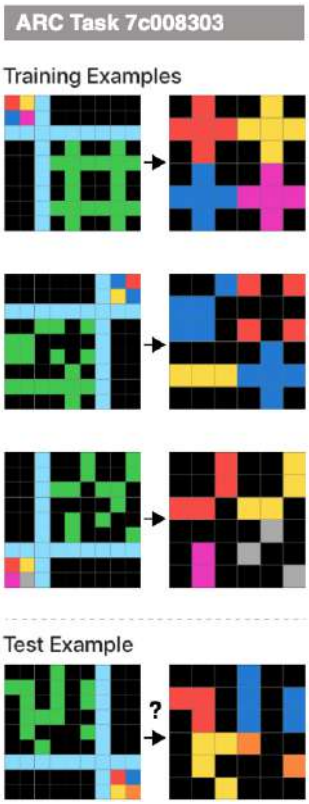


Figure 2.5

mental trajectory.

Object persistence is the ability to recognize objects and manipulate them without a change in orientation or context causing the identification of that object to disappear. It is, first and foremost, about determining whether an object exists within the image or not, but also about understanding whether it exists necessarily or contingently. Referring back to the example in figure 2.5, we observe that the presence of the yellow squares, amidst the green shapes, is necessary. In contrast, the presence of the green squares is contingent. There is no rule governing their arrangement. It also follows that it is impossible to have a yellow square outside of a closed green shape. These three modalities—existence/non-existence, necessity/contingency, and possibility/impossibility—are precisely addressed by Kant's category of Modality. This allows us to distinguish between problematic judgements (it is possible for a square to be green), assertoric judgements (these squares are green), and apodictic judgements (a square is yellow only if it is within a closed green shape).

Goal orientation is the ability to ascribe intentions or goals to certain objects. Solving the Raven's matrix in example 2.5 is easier if you think that the "goal" of the yellow pixels is to fill the closed spaces. You might think of them as a herd searching for the shelter of an enclosure at nightfall. This ability does not strictly correspond to the categories identified by Kant. Nevertheless, it seems to belong to the category of Relation. This category indeed prescribes causal relationships between objects. It allows for the differentiation of categorical judgements (the herd is seeking shelter!), hypothetical judgements (if the herd wants to survive, it must find an enclosure), and disjunctive judgements (either the animal finds an enclosure, or it disappears). I recognize that this parallel I draw between goal orientation and Kant's category of relation can only be purely analogical and resists a thorough analysis. The reason for this is that Kant is not concerned with intentionality, but with the different forms of judgement—considered objectively.

Finally, *geometry and topology* are, according to François Chollet, the ability to perform spatial manipulations on objects, such as rotations, orthogonal symmetries, translations, deformations, repetitions, associations, etc. This requires systematizing the qualities of objects and the relationships between them. One must, in particular, understand situations where a change in the image is either accidental or necessary. When a 45-degree rotation occurs, for example, it is necessary for certain points to move, but not in any arbitrary way. This ingredient I discuss here straddles Kant's categories of Quality and Relation. Quality allows us to assign intrinsic properties to objects, such as being green or yellow. Relation, on the other hand, concerns not intrinsic qualities, but the relations that these objects have with each other: spatial distance, the presence of symmetry, and so on.

The analysis I have just made of François Chollet's criteria seems to sug-

gest that, in order to achieve strong artificial intelligence, it is necessary to endow algorithms with a certain ability to judge. This ability relies on a set of intrinsic structures of thought. In other words, our judgements follow a set of conceptual distinctions such as necessity, contingency, existence, non-existence, possibility, impossibility, and causality. These *a priori* concepts are the building blocks of reasoning; they are what allow any logical proposition to be expressed or understood. It is therefore crucial to ask what the conditions are for a machine to understand and use these concepts. I will address this important issue in chapter 3.4.

At this point, I would like to return briefly to Kant's thought. In the approach of the *Critique of Pure Reason*, defining the logical categories of understanding is not enough; one must also analyze what guarantees their validity, both from an objective and subjective standpoint. Let me clarify this distinction. Objectively, it is necessary to provide a guarantee that there are indeed objects to which the concepts we have defined, derived from pure logic, apply. In other words, it must be shown that these are not empty categories.

From a subjective perspective, we must also explain the internal workings of understanding; that is, we must explain what makes it possible for consciousness to effectively synthesize the diverse data of sensibility and subsume them under a category. In fact, this second question is not strictly necessary for the critique of reason itself, but it provides very valuable insight. Kant himself says: "Although this examination is of great importance relative to my main goal, it is not essentially part of it, for the crucial question is always to know what understanding and reason, free from all experience, can know, and to what extent they can push their knowledge, and not how the very faculty of thought is possible[36]."

2.4.2 Objective Deduction

It could be the case that no sensible content corresponds to the categories thus determined; that phenomena follow each other in such a chaotic manner that nothing within them provides a rule corresponding, for example, to the *a priori* principle of cause and effect. In such a case, phenomena would present objects to our intuition, but would not be intelligible, and it would follow that the categories of understanding would be nothing but pure, empty forms. "Explaining how concepts can relate *a priori* to objects, that is what I call the transcendental deduction of these concepts[37]."

This objection, fortunately, dissipates like smoke as soon as one takes the trouble to remember that, for Kant, it is not the understanding that must adjust to the objects, but rather the objects that adjust to the understanding. Subtlety and wordplay? Far from it. What Kant asserts is that

[36] *Ibidem*, Preface to the First Edition, pp. 11-12.
[37] *Ibidem*, Transcendental Analytic, Chapter II, Section 1, §13, p. 148.

the understanding, through its concepts, is the creator of experience. The objects, in terms of their phenomenal appearances, therefore adapt to the *a priori* concepts within us. Consequently, it is impossible for these concepts to find no object. Or in other words, these *a priori* principles are always the principles of "true" phenomena, although these phenomena teach us nothing about the thing-in-itself.

An important consequence that follows from all of this, and which significantly restricts the use of understanding, is that the categories always unfold within experience; if not a concrete and actual experience, then at least a *possible* experience. The term "possible" is of paramount importance here. In the case of *a posteriori* judgements, such as those regarding concrete objects through an actual experimental setup, there is no doubt that the categories have objectivity. In the case of synthetic *a priori* judgements, however—the very ones Kant attempts to prove the existence of—experience cannot touch actual objects presently given to our senses. Nevertheless, the terms of the judgement, such as the ideas of "heavy body" or "metal," can only be thought within the framework of a possible experience. The best example of this type of reasoning in physics is thought experiments. They are used to test the laws of nature and understand their coherence, or conversely, to invalidate them. Outside of this experience, the categories are without object.

Kant heavily emphasizes this aspect. Certainly, the categories have an a priori existence, and certainly, they allow for the definition of fields of pure knowledge, such as those of mathematics or physics. But the fact that they are valid in an a priori manner does not prevent knowledge from needing, in order to be made effective, to be accompanied by a possible experience. Mathematical ideas do not escape this rule. Although there may not exist in the world a single object perfectly conforming to the idea of the triangle, this idea still derives its meaning from the possibility of being applied to sensible content.

Finally, note that the laws of judgement, which are the categories, are the conditions for the possibility of experience. It is not experience that makes concepts possible, but, on the contrary, it is concepts that make experience possible. Nevertheless, the concept must have content; it must be directed towards an intuition that is sensible. As such, thought is not solipsism, it is not pure immanence; instead, it refers to a beyond of pure thought. For Kant, real knowledge is synthetic knowledge. This is probably why he eloquently asserts that his philosophy avoids both the pitfall of empiricism and that of idealism. Although the noumena are beyond access, although the subject gives itself its own phenomenal laws, representations are not mere illusions. There is truly an encounter with otherness, with a matter that is not contained in an a priori manner within the subject. However, this otherness is given within the modalities set by the pure intuitions of space and time, then informed by the categories of understanding.

2.4.3 Subjective Deduction

Kant addresses here the question of the constitution of knowledge: how is the act of synthesis, which occurs within the understanding, possible? He takes the example of a man entering a magnificent palace. His senses are struck by the brilliance of the stained glass windows, by the refined elegance of the furniture, by the grace of the colonnades, and by many other details, each vying for his attention. Yet, his consciousness is not torn between all these diverse impressions. Our man is able to identify all these elements, differentiate them from one another, and isolate them from their surroundings; yet at the same time, he retains the certainty of the unity of his own consciousness. It is the same man who is conscious of the stained glass window, the furniture, the colonnade, and not a plurality of heterogeneous consciousnesses.

In other words, consciousness is composed of a multiplicity of sensible perceptions, but it remains invariably one. This is therefore a synthetic *a priori* unity. Of course, this unity is the everyday reality of our sensible experience. It has become so natural to us that it may seem strange to ask where it comes from. Yet this is precisely the fundamental question posed by Kant, when he asks "what contains the principle of the unity of different concepts within judgements and, consequently, of the unity of the understanding itself[38]."

In response to this question, Kant immediately points out that the unified consciousness of the self is not to be sought in a category, such as the category of unity, nor in any category superior to those already given. Indeed, the category merely represents a logical function of judgement. As such, it already presupposes the unity of the concept, or in other words, it cannot explain how the categories are applied. Proceeding in this way would only shift the question. It would be akin to trying to account for the operation of an unknown machine using that very same machine.

Where, then, should one seek the principle that underlies the unity of our sensible representations? The answer: in the Cartesian "I think," which brings the diversity of sensible perceptions back to a single and unified self-consciousness. "The *I think* must be able to accompany all my representations; for otherwise, there would be something represented in me that could not be thought, which amounts to saying either that the representation would be impossible or at least that it would be nothing for me[39]."

Let us clarify this idea. The diverse representations given in an intuition, for example in the intuition of a palace or a temple, are all together my representations only because they all belong to the same consciousness. Consciousness, therefore, is not scattered across each of these representations, but it is fully present in each one of them. This principle is what

[38] *Ibidem*, Transcendental Analytic, Chapter II, Section 2, §15, p.160.
[39] *Ibidem*, Transcendental Analytic, Chapter II, Section 2, §16, p.160.

Kant calls *transcendental apperception*. It is crucial in Kant's system. It is to the understanding what intuition is to sensibility: namely, a form that makes the exercise of the understanding possible *a priori*. "The supreme principle of the possibility of any intuition, in relation to sensibility, was, according to transcendental aesthetics, that everything it contains in diversity was subject to the formal conditions of space and time. The supreme principle of this same possibility, in relation to the understanding, is that everything diverse in intuition is subject to the conditions of the originally synthetic unity of apperception[40]."

This general form in turn subdivides into a number of particular forms corresponding to the different logical functions of judgement, which Kant called the categories. Transcendental apperception is therefore the origin of all the concepts of the understanding, and represents its higher principle of operation.

2.4.4 Kantian Schematism

As long as they have not been subsumed under a category, the sensible data as such cannot be *perceived* by consciousness. They are only confused perceptions, without form or distinguishable unity. This is why, now that the objective and subjective possibilities of the categories have been verified, it is still necessary to specify how synthesis operates concretely; the mechanism must be shown.

To this end, Kant identifies the presence of two ingredients: imagination and pure intuition of time. Why time and not space? Because time is common to all phenomena, both those occurring within us and those external to us, while space is merely an *a priori* form of phenomena responding to external sense. Now, in order for synthesis to take place, it is essential to link the category, which is *a priori*, with what the phenomenon itself has *a priori* within it. Imagination, for its part, has a *henological* function. It produces schemas and images that mediate between the synthetic unity of consciousness and the multiplicity of sensible representations.

What is a schema, and how is it different from an image? What is their relationship? The schema has no visual component. It is neither a representation nor a concept but a method of constructing the sensible object within intuition. Let us explain. The concept of a triangle has an extension that includes an infinity of possible triangles: obtuse-angled, acute-angled, isosceles, right-angled, etc. The cognitive function that allows us to associate the image of a triangle with its mathematical concept proceeds through a mental recognition mechanism that verifies whether this image of the triangle indeed corresponds to the definition we hold of it. We mentally trace the contours of the geometric figure, ensuring it indeed has three sides. This

[40] *Ibidem*, Transcendental Analytic, Chapter II, Section 2, §17, p.164.

process takes place in time and is entirely *a priori*: even in the absence of a "real" triangle, it is always possible to construct one mentally. The existence of a mathematical object, from this point of view, is constructive. What exists is what can be defined through a numerical or geometric process involving a finite number of operations.

Once the image of the triangle has been identified using the schema, a mental image of the object forms. This blurred image is neither precise enough to correspond to a real sensible triangle nor abstract enough to represent the concept of the triangle in its full generality. It often includes properties not belonging to the definition of a triangle, such as colours, a certain thickness, or a specific size. Nevertheless, it is this mental image, derived from the schema, that will serve as the intermediary between the sensible object and the concept, enabling the synthesis.

This theory of the schema changes the value given to imagination. Plato viewed imagination as a corrupted faculty, and for two reasons: not only did it pertain to the sensible, but far from providing us with the representation of the object itself, it gave us only a pale copy[41] ! It is for this reason, no doubt, that imagination was associated with the most vulgar level of knowledge: *eikasia*. It was so deeply despised by the philosophical tradition that it was alternately seen as a source of errors or as a source of temptation. Religion and metaphysics were its most faithful detractors. Yet, with Kant, it is revalorized. Imagination is no longer merely the capacity to produce images, but it is also and especially what makes possible the unfolding of intuition and the sensible representation of our concepts. It is no coincidence, therefore, that Kant devotes an entire chapter to it in his *Anthropologie*[42], and gives it an important place in aesthetic judgement[43].

Despite the role played by imagination, synthesis remains on the side of the intelligible. The first reason for this, and probably the most obvious, is the fundamental place occupied by the categories. Their function is to extract conceptual units from the sensible, by performing a series of divisions and associations. This discursive function is analogous to that of *dianoia* in Plato's system. The second reason synthesis remains close to the intelligible is that it incorporates schemas. These are neither images nor forms of reasoning but a kind of contemplation of concepts. However, this contemplation is not immediate, as Plato's *noesis* was. It is a process, a mental

[41] Imagination can be found in Plato's dialogues under the Greek terms *phantasma* and *phantasia*. The first term refers to *doxas* represented independently of any sensation (productions of the imagination), while *phantasia* is more properly understood as a distortion of sensible perception. For more details on this topic, see Collette [**?**], *Phantasia and Phantasma in Plato*. Note that some authors (Baccou, Cazeaux, Karsenti/Prélorentzos) also translate "Eikasia" as imagination (see section 1.2, *The Analogy of the Line*).

[42] cf. Kant [**?**], *Anthropology from a Pragmatic Point of View*. Chapter §xxviii is entitled *On Imagination*, and Chapter §xxx, *On Productive Imagination*.

[43] Unlike Plato, who admitted only what contributed to moral education and truth in art.

construction of sensible objects. In short, Kant re-inscribes the intelligible within the sensible, but without erasing its noble contours. One could say that with him, the heavens of Ideas have fallen on our heads. Imagination has been sublimated; intuition has become flesh.

2.5 Reason: The Man Who Thinks He Thinks

As long as the categories of the understanding are applied to sensible intuitions, knowledge remains founded *on* experience *through* the *a priori* principles of the mind. But when pure thought, independently of any content, rises to transcendental ideas according to the deductive thread that logic permits, it soon loses itself in abstract and antinomic speculations. If the understanding seeks to bring the unity of sensation back to certain rules, reason, for its part, seeks to bring the unity of these rules back to certain principles.

Kant distinguishes three transcendental ideas: *the soul*, *the world*, and *God*, studied respectively by rational *psychology*, *cosmology*, and *theology*. These ideas are derived from the three forms of reasoning (or more precisely, the prosyllogistic series of reasoning) that are *categorical*, *hypothetical*, and *disjunctive* reasoning. Kant analyzes the sophisms arising from each of these three types of reasoning in order to circumscribe the limits of the validity of reason. We will not delve into the details of each of these discussions, as this would take us too far from our subject. We will merely exemplify the approach in the case of the soul, which, of the three antinomies, is the easiest to relate to the Socratic tradition.

According to rational psychology, the concept of the soul can be derived by reasoning from the "I think", without needing to refer to experience or sensible intuition. The main properties of the soul are directly derived from the table of categories, by applying the pure concepts of the understanding to the ego cogito [44]:

1
The soul is
a substance

2
Simple in
its quality

3
Identical
over time

4
Related to objects
possible in space

[44] KANT [99], *Critique of Pure Reason*, trans. by BARNI, Transcendental Dialectic, The Paralogisms of Pure Reason, p. 343

From the first of these predicates, rational psychology derives the idea of the immateriality of the soul; from the second, the idea of its incorruptibility; from the third, the idea of its personality. When taken together, these three lead to the idea of the soul as spiritual and immortal. Finally, the fourth predicate gives the idea of the soul as a vital principle, to be compared with the concept of anima in Greek thought.

However, for Kant, these reasonings are mere paralogisms. The "I think" does not allow me to take myself as an object of knowledge, and for two reasons: 1) on the one hand, there is no sensible intuition of myself in pure thought, and consequently, the ideas derived from it are mere empty forms; 2) on the other hand, I can only know myself in the "I think" as the unity of perception, not as a substance. It is therefore a mistake to imagine that one can deduce anything about the being of the self from the ego cogito of pure thought.

Two remarks can be drawn from this analysis[45]. First, we observe that Kant's theory does not permit a philosophy of subjectivity as such. The subject is not capable of taking itself as an object of knowledge, or when it does, it presents itself as an appearance (in the case of sensibility), as a mere unity of consciousness (in the case of understanding), or as an illusion (in the case of reason). In no case can we speak of a reflexive subject, capable of giving itself its own concept and thereby entering into a dynamic of self-overcoming. On the contrary, the Kantian system is steeped in statism and formalism. The ancient gnwti seauton ("know thyself") advocated by Socrates can no longer apply. From then on, the pursuit of knowledge no longer implies the need for the subject to transform itself in its very being. Far from the initiation rites practiced in the corridors of Pythagorean and Orphic sects, far from the education of ideas and the exit from Plato's cave, the Kantian system invites the subject to bring the object of their knowledge into a laboratory. It is no longer about changing in relation to the nature before us, whose mysteries we try to unravel, but about making the object of experience identical to oneself, subjecting it to our principles. We will find nothing in it other than what the understanding itself has placed there. This logic of the identical marks the exit from spiritualities and the entry into rational knowledge and the philosophy of the Same.

Another major remark is that, unlike Plato, the innate structures of the human mind are no longer defined on the basis of a metaphysical model. For both Plato and Kant, the mind possesses knowledge prior to any experience, knowledge that is already in it and that it merely (re)discovers. In Plato's view, this is made possible because the soul, being immortal, has already contemplated the Ideas, but has lost the memory of them upon its incarnation. Knowledge, therefore, consists of reminiscence; education is a maieutic

[45] These two remarks are mine, as well as the analysis and summaries of Kant. I have based myself solely on the reading of the original sources.

process. Learning is giving birth to forgotten things. This is why, for Plato, Reason is a contemplation (noèsis) of what is always already outside of oneself. In contrast, for Kant, knowledge must dispense with any metaphysical foundation. Pure Reason leads only to error and confusion. If there are indeed a priori truths, it is because the entire reality appears in accordance with our inner dispositions, which are our pure intuitions and categories. The appropriate tool is no longer to be found in contemplation or in the free exercise of pure thought, but in the understanding. This faculty works in a manner similar to Plato's dianoia, reasoning always from hypotheses and possible experiences, but with the help of sensible intuitions. For Kant, it is therefore about intuiting outside of oneself what is always already inside oneself.

2.6 Beauty Without Truth

What is the new conception of judgement that emerges from the Kantian revolution? One may recall that, for Plato, the True, the Beautiful, and the Good represented three divine aspects, complementary facets of a fundamental essence—a unity to which Socrates himself compared the Sun. To contemplate one of these Ideas wholly and authentically is akin to peering beneath the surface of a lake. As long as one remains above the water, one mistakenly perceives scattered rocks, tree stumps swallowed by silt, solitary waterweeds. Yet, when one plunges beneath the surface, one realises that these rocks, these stumps, these grasses constitute one single whole. It is impossible to fix one's eyes on any single fragment without glimpsing the network of relations that binds them. In Plato, truth, moral goodness, and beauty are entwined so tightly, so vividly, that no beautiful action could fail to be equally true and good. Is this not, indeed, why the ancient Athenians, with fervent conviction, maintained that the good life was also, and above all, a beautiful life?

It is precisely this immutable solidarity of the Ideas that Kant sets out to challenge in his various critiques of pure reason, of practical reason and judgement, and he does so in three ways. Firstly, (1) he removes from speculative Reason the prerogative of truth; in addition, (2) he bestows upon Reason the authority to distinguish good from evil; and finally, (3) he regards beauty as the free play of the sensible imagination. The first of these three points is purely *critical* and *restrictive*. As we recall, Kant demonstrates that the great questions concerning God's existence, the finitude of the Universe, or the immortality of the soul are undecidable. The very concepts mobilised in these inquiries are inadequate, such that it becomes possible to prove both a claim and its contrary. Scientific truth, for Kant, does not depend solely on analytic judgement, but on one's experience of the world. This experience, however, is shaped by our concepts, by the a priori structures

of our mind, so that scientific truths are, above all, regularities of human perception.

In denying Reason the capacity to apprehend the True, and in founding the truth of the sciences on the a priori forms of Sensibility and Understanding, Kant's critique profoundly unsettles the entire scaffold of metaphysics. It is now futile to aspire to that mystical, nearly esoteric *noesis*, which underpinned the Platonic school. The goal of philosophy is no longer to transform oneself from within or to elevate one's soul towards the realm of intelligible being. From now forth, it is to critique, by means of the transcendental method, the illusions of metaphysics.

Though Kant denies Reason the power to inform us regarding speculative metaphysical questions, he does not strip it of the privilege of judging good and evil. This is point (2). In the *Critique of Practical Reason* and the *Groundwork of the Metaphysics of Morals*, Kant endeavours to elucidate the a priori laws of morality. These are founded upon the principle that the human being is an end in themselves. If Nature has endowed us with this strange capacity for reasoning, when it could just as well have directed us by means of the stick and carrot of pleasure and pain, it follows that reason must serve some essential purpose. Each individual's capacity for thought—that truly extraordinary marvel—must be prized and respected.

It is this capacity for thought that grounds our human dignity, that infuses life with its salt and its sting, corroborating the mysterious nature of humanity itself. For as long as a being is capable of logical, universal thought, they cannot deny that any other being so endowed deserves, equally, to be respected as an end in themselves. Thus, the exercise of morality is rendered possible by the alignment of our actions with this fundamental imperative[46]:

> *"Act in such a way that the maxim of your action could be willed as a universal law; act so that you always treat humanity, whether in yourself or in others, as an end and never merely as a means; act as though you were both the legislator and the subject in the republic of free and rational wills."*

Thus, morality is founded not upon compassion or social interest, but upon disinterested duty. This duty consists in respecting the autonomy of reason, and, through it, all rational beings. One may therefore declare that, for Kant, the good is deontological. It rests upon laws that reason must enable us to elucidate. Moreover, such reasoning is not a matter of contemplation or internal transformation. It pertains more to discipline, to methodical rectitude. There is something austere, even disembodied, in this vision of the Good—an idea which would have found little echo among

[46] The passage below compresses the three Kantian imperatives, as formulated in the *Groundwork of the Metaphysics of Morals*.

the ancient Athenians, steeped as they were in the mysteries of Eleusis, the Bacchanals, or the rites of Pythagoras. Does this mean Kant is utterly indifferent to the Beautiful? That he ceases to discern in it a privileged link, a gateway, a tutelage leading to higher truth?

On the contrary, in many of his works, Kant defends the view that, although morality originates in Reason alone, the moral sentiment partakes of its force and grandeur. The emotion stirred by heroism—especially sacrifice—or virtue exercised without regard for self-interest, is, in his eyes, a constant of human nature.[47] Far from being a purely Teutonic exaltation, mingled with patriotism, Kant asserts that this emotion is universal, the constitutive foundation, the fertile soil of early moral development. Thus, the Beautiful, though distinguished from the Good, is not thereby forsaken. It retains a central role in human judgement.

This role, the third point, is not however what Plato imagined. For the latter, beauty was a kind of Ariadne's thread, laid out in the labyrinth of life for the sake of the wise, to guide them step by step towards the path of truth and supreme goodness. Thus, the young, dazzled as they are still by the vividness of the senses and pleasure, first pursue the impulses of Eros. They learn to love beautiful bodies, to embrace and possess them, before turning towards beauty without form, face, or matter. They ascend, step by step, toward intellectual contemplation, starting with illusion and belief, and arriving finally at pure reason.

By contrast, for Kant, beauty is not an end in itself. He devotes his third major work, the *Critique of Judgement*, to the question of the Beautiful. Here, he contends that aesthetic experience is the disinterested play of imagination and sensibility. By *disinterested*, one must understand that what is beautiful is devoid of utility and not synonymous with mere pleasure. That which springs from amorous passion or carnal desire is beautiful only insofar as it brings pleasure. Indeed, that very body we once judged beautiful, should it withhold its favours, might suddenly appear detestable. This inversion—so familiar to poets—that transforms adoration into hatred, or hatred into passion, is proof that what is at stake is not beauty itself, but ourselves: our interest, our pleasure, our future.

In other words, Kant stipulates from the outset that beauty is not an end, and as such, has nothing in common with reason. Hence the difficulty, even impossibility, of expressing reasonably why some things are beautiful

[47] See the *Groundwork of the Metaphysics of Morals*, where Kant writes: "For the most common observation shows that if one presents an act of honesty, unblemished by any view to personal interest, whether in this world or in the next, performed with a steadfast soul even amidst the strongest temptations of want or seduction by some advantage, it surpasses and outshines any like act which, even in the slightest degree, has been affected by some alien motive; it uplifts the soul and excites the desire to be able to do likewise. Even children of middle age feel this impression, and it should never be presented to them in any other way." (p.26)

and others, not. Attempts to do so lose sight of beauty's essence. One invokes explanations: forms are pleasing, colours are gentle, melodies are delicate, or the artist behind the work exhibits such-and-such quality.

But one must take care not to jump from Scylla to Charybdis: though the judgement of taste is non-rational, it has a sort of universalizability. I say universalizable, not universal, for tastes and colours, as the old adage goes, can never be identical for everyone. Yet, the one who judges something beautiful invariably assumes that beauty is an objective property of the thing. Although the sentiment of beauty arises from within, we speak as though all ought necessarily to feel the same way. Thus, there exists, in aesthetic experience, a tendency towards generalisation—a tendency rooted not, as in morality, in the universality of reason.

What, then, is the ingredient unique to the judgement of taste, prompting this inclination towards universalisation? Kant identifies two: imagination and the categories of the understanding. Yet these faculties are not employed as they are in forming concepts. In that instance, the category imposes logical structure on phenomena, separating objects, applying causal laws, distinguishing possibility from necessity. The imagination, for its part, plays a crucial role in schematisation—enabling us to picture an abstract concept, such as a triangle, by means of a vague, coloured, archetype form appearing against the blankness of a page.

Yet in the case of the judgement of taste, there are no concepts involved. Here Kant is unequivocal. It is not the image of the apple in the midst of the Garden of Eden, delicately held aloft in outstretched hands, that is to be termed beautiful. Rather, it is those artistic portraits that distort concepts, that idealise them, that subvert ordinary conventions, which are considered beautiful. Within the human mind, there exists a profoundly creative process, one that finds its pinnacle in aesthetic experience. Far from remaining a passive receptacle, taking in the imprint of external beauty, the subject, astonished by the beautiful, actively participates in the experience of wonder.

Alongside this first kind of aesthetic experience—free and creative—there stands as well the experience of the sublime. This, once again, is encountered when what is represented no longer matches its concept, but in a very particular way: when what is being expressed is too vast, too intense, too immense to be circumscribed within any representation; when a landscape, a tempest, a feeling of heroism appears to defy the careful boundaries of our concepts, then the mind is seized by its own infirmity. It succumbs to the vastness of the world; it trembles, throbs, and vibrates in harmony with that which surpasses it. At the very core of beauty, in its noblest and most overwhelming aspect, lies the idea of the infinite.

To summarise Kant's conception of beauty, as it emerges from his critique, we might say that beauty has become disentangled from its ancient ties to the good and the true. To each of the three Kantian fac-

ulties—Sensibility, Understanding, and Reason—corresponds one of Plato's three ideas. The Good becomes a purely rational, logical, universal law. The True is given through sensible experience and is structured by the categories of Understanding. Beauty, for its part, is the beyond of representation, the autonomy of the sensible world, the stunned confrontation with infinity that defies the limits of understanding.

In the second part of this thesis, I shall attempt to show how modern thought, embodied by the functioning of machines, continues this deconstruction of the Platonic tripartition. With the advent of calculation and logical formalisms, and the binary languages underpinning computer science, rationality becomes the only criterion of truth. What is valid must be subject to logical, automatic deduction. Observation, intuition, and instinct, unless pre-ordered by a methodological, scientific, and computational framework, are met with suspicion.

This movement is what I shall call the forgetting of the Understanding. It consists in flattening the complexity of the human mind—and the world that manifests itself through it. From henceforth, one sees only two levels of thought, two faculties: sensation and rationality; synthetic a posteriori judgements and analytic judgements. The judgement of taste becomes a private affair, with no claim to universality. Truth is reduced to patterns of correlation. The good, nothing more than an abstract rule, is cut adrift from the human heart. There is no longer any connection, even a partial one, between the beautiful, the good, and the true.

This forgetting of the Understanding, this abandonment of beauty, this relativism of truth, explains in part the quasi-sacred trust granted to algorithms. They are the products of a historical turning point, of an ideological purification begun with the advent of modernity, which consists in stripping away all that remains of the transcendent and contemplative in human experience.

The second part, entitled *Revision of the Concept of Judgement*, will therefore aim to explain how we have come to forget beauty, and how this forgetfulness affects the judgements made by the machines we create. However, before launching a systematic critique of modern algorithmic rationality, I wish to deploy the philosophical instruments developed thus far, through Kant and Plato, to reflect further on our philosophical trajectory: in light of what has been established, can it truly be said that a machine is capable of judgement?

The next chapter will distinguish several types of judgement. I shall attempt to ascertain whether algorithms can render rational judgements, judgements based on experience, even synthetic a priori judgements. Such preliminary considerations will enable us to ask whether a machine possesses any intuition of space and time, whether it has concepts, imagination, or some subjective sense of being a subject—a 'self' with a point of view on the world. These considerations will serve as the point of departure for the

reflections that follow. But first, a few methodological remarks are in order regarding the importance I have attached, throughout this thesis, to Kant's first critique.

2.6.1 The Importance of the First Critique

A conscientious reader might ask why, in this thesis, I have elected to focus on the *Critique of Pure Reason*, according greater weight to it than to the second critique (on practical reason) or the third (on aesthetic judgement).

Beyond the subtle arbitrariness, woven with subjective preference, that inevitably hovers over one's choice of foundational literature, I wish here to highlight some objective considerations that make the first critique the ideal subject for my approach.

To this end, perhaps it is worth recalling the strategic objective of this analysis of Kantian thought. Far from attempting a merely descriptive archaeology of his philosophy, in the manner of a scholar or scholastic eager to parade their erudition, my chief aim is this: to show that there is a continuity, a line of descent if you will, between the concept of judgement as found in Plato and the concept found in Kant.

More precisely, I sought to show that this dual-level theory in Plato, opposing sensation and reason, gains new depth through the insertion, between these two faculties, of Understanding. This marks the inaugural movement orchestrated by Kant's *Critique of Pure Reason*, a movement which, furthermore, is carried out in full awareness of its radical nature.

When discussing the concepts of Understanding, I pointed out that Kant refers explicitly to "the dove that cleaves the air with no resistance"[48] (see the section "The Dove-Cote Metaphor"). In these words, he invokes Plato's dovecote metaphor, whereby judgement is said to be true or false ideally, without reference to sensory content.

For Kant, it was crucial, at the moment of introducing the judgement of understanding, to show that there exists something between pure reason as envisioned by Plato—concerned with the unconditional—and sensation itself. If this reference to Plato seems uncertain, consider, too, that Kant repeatedly cites the example "5 + 7 = 12" (Theaetetus, 196a), which opens the second preface of the *Critique of Pure Reason* to illustrate synthetic a priori judgement. This example, once again, is drawn from Plato's account of the dovecote in the Theaetetus.

This, I believe, justifies my choice to focus on the *Critique of Pure Reason*, and not on Kant's other works. It is by no means the only way to ground a critique of artificial intelligence—far from it! The dichotomy between determining and reflecting judgement, which Kant elaborates in the *Critique of Judgement*, may prove of interest. Nevertheless, I am persuaded

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

that these concepts are less relevant than the three-level system of sensibility, understanding, and reason.

Determining judgement, in modern parlance, refers to the deductive approach, where a judgement is constructed from a universal principle. In the syllogism "All men are mortal; Socrates is a man; therefore, Socrates is mortal," we move from a universal truth to a particular one.

Reflective judgement, on the other hand, pertains to induction. Far from proceeding from reason's categories, it concentrates on phenomena, seeking to study them, to let them reveal themselves, and thus to determine more general truths. It operates, so to speak, in the dark, without prior determinations.

At first glance, these distinctions might map—albeit imperfectly—onto the top-down and bottom-up approaches I outlined in the introduction to this thesis. Top-down strategies, prominent in the early days of artificial intelligence, consist in defining a set of general principles and proceeding towards conclusions from them. The archetypal example is the so-called expert system, which I shall discuss in Section **??**.

Bottom-up strategies, by contrast, begin with data from experience and refine their decisions through interaction with it. They embody a form of learning with minimal or no human supervision: the ambition of second-generation artificial intelligence, of deep learning and neural networks—of which ChatGPT is the most spectacular outcome.

One might criticise this, admittedly hasty, parallel I have drawn between the bottom-up approach and Kant's reflective judgement—especially when considering concepts such as freedom or aesthetic judgement, which are indispensable for Kant, yet seem entirely subordinated to computational processes within contemporary algorithms. Still, this opposition between reflective and determining judgement only recognises two levels, not three. It does not, therefore, fully serve the purpose of my project.

I hope that these remarks suffice to show that the construction of this thesis—though it certainly does not wholly escape arbitrariness—follows an inner logic that would not have been so easily justified by an interpretative focus on either the second or the third of Kant's critiques.

What to take away from this chapter

Kant critiques the conceptions of Plato and develops a theory of judgement based on three faculties:

1. *Sensibility* is the ability to make judgements of experience; this synthesis, however, relies on prior forms of human thought, namely the intuition of time and the intuition of space.

2. *Reason* is the ability to make rational judgements. This analysis is not based on any data from experience; it merely thinks through the purely logical articulations of an idea.

3. *Understanding*, for its part, is the ability to make synthetic *a priori* judgements. This somewhat particular judgement is a revision of Plato's ungrounded judgement. On the one hand, it rests on the very structures of experience (the intuitions of time and space), but on the other hand, it touches upon universal concepts (the twelve categories of Understanding).

We thus understand that Understanding is an essential link between synthesis and analysis; between experience and logic. I have also suggested that Understanding could be the missing ingredient in the computer. When faced with François Chollet's ARC exercises, algorithms indeed struggle to generalize and abstract their observations.
In the next chapter, I will discuss in more detail the structural differences between human and current algorithmsin a Kantian framework.

Chapter 3

Artificial Judgement: Can the Machine Judge?

We're like the man who climbed on a chair and declared he was a little closer to the moon.

Hubert Dreyfus

It is now time to ask what Kantian theory can teach us about artificial intelligence. At the beginning of our discussion on Kant, we encountered a stumbling block, which we attempted to navigate around skillfully. This stumbling block is François Chollet's ARC tests, where current algorithms yield results that are more than mediocre (see figure 3.1).

These exercises are similar to Raven's matrices, but less standardized and more creative. They involve generalizing certain situations based on only two or three examples. It was evident that humans find it easy to give the correct answer, but not so for algorithms. It seems as though, for us, the patterns of generalization are embedded in our perceptual structures. These structures don't seem to depend much on experience; although they develop in contact with the world, they are innately present in us. They allow us to grasp spatial symmetries, rotations, and the abstract thought operations found in Euclidean geometry theorems, such as when we say that the sum of angles equals 180 degrees.

The problem that arises is whether the abstract thinking required to solve the ARC problems could not be akin to the synthetic *a priori* judgements I introduced in the previous chapter? If so, we would gain an interesting insight into what differentiates human judgement from artificial intelligence. The synthetic *a priori* judgement, in fact, fundamentally rests on pure forms. According to Kant's theory, they exist within us prior to any experience: they are the intuitions of space and time (chapter 2.3), as well as the categories of the understanding (chapter 2.4).

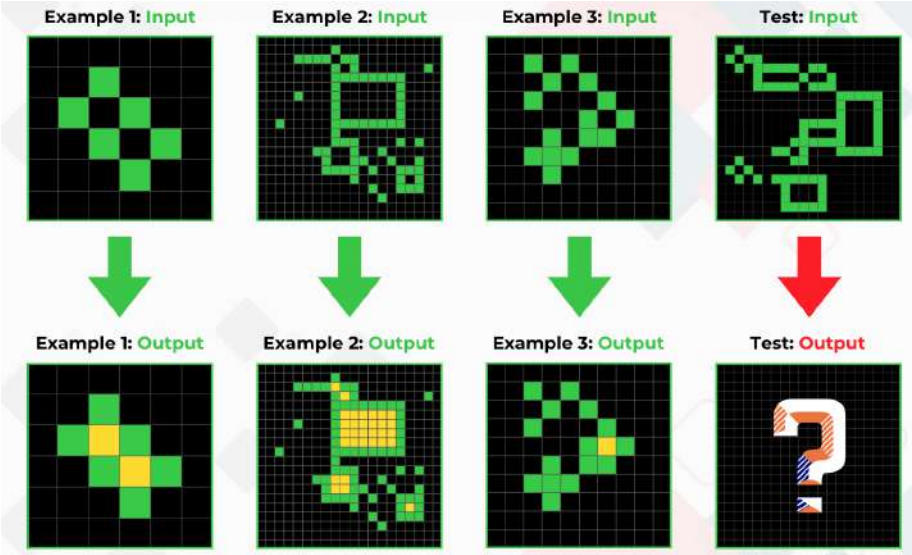


Figure 3.1: Here is an example of an ARC test. Given an input, the machine must produce the correct output. It receives three examples to understand what it should do (source: `https://lab42.global/arc/`).

In contrast to the sensors of a robot, which merely capture environmental data passively, these *a priori* forms are not mere receptacles. They are structures, filters, that *inform* the world, occasionally even *distorting* it. In other words, we are equipped with geometric concepts that shape the way we perceive the world, enabling us to isolate objects, abstract properties, and make logical conclusions, even in the face of problems on which we have not been specifically trained (or very little).

The question I would like to raise in the remainder of my thesis is whether, within the human mind, it is not the sensory data that adapts to our *a priori* concepts, while within algorithms, it is the concepts that struggle to adapt to the data? This idea I am suggesting would explain why we find Raven's matrices easier to comprehend. These matrices are designed to match our thought structures. When presented to us, we expect to see symmetries, translations, squares, and rectangles. The machine, on the other hand, lacking any a priori, must attempt to grasp these transformations solely based on examples. Since these examples are insufficient, it produces erroneous responses.

To illustrate the presence, in humans, of structural geometric concepts, I would like to focus on certain optical illusions, particularly Kanizsa's illusion (see figure 3.2). In this illusion, black discs are cut in such a way as to suggest, in the blank space of the page, the presence of a square or triangle. These shapes, which we feel we are seeing, are a pure product of the imagination. It's as if our mind is drawing the contours of a non-existent object.

Figure 3.2: In Kanizsa's illusions, a geometric shape appears in the blank space of the page. From a Kantian perspective, this illustrates that it is the sensible world that adapts to our categories, not the other way around.

These imaginary edges we observe echo the Kantian theory of schematism. In this theory, let us remember, the human mind is not a mere passive receptacle. It plays an active role in perception. The imagination is mobilized to cut up reality and construct the contours of the concepts with which we are most familiar. This is why Kant says, "The schema is always by itself only a product of the imagination [...] serving to provide an image to a concept"[1].

When the Kanizsa illusion is presented to software such as ChatGPT, the results are mixed. The algorithm is indeed capable of identifying the presence of the triangle and the square in the original scenarios, but if a new configuration is presented, with a random perturbation of the original schema — one that does not exist on the internet — a very different response is obtained. The discussion below was obtained by submitting the image on the next page (figure 3.3) to ChatGPT[2] (version o1, November 2024):

> HUMAN: A geometric figure appears in the blank space of the page. What is it?

> ChatGPT: The geometric figure that appears in the blank space of the page is a six-pointed star. It is formed by the negative space between the more amorphous black shapes.

[1] KANT, *Critique of Pure Reason*, Volume I, On the Schematism of the Pure Concepts of the Understanding, Chapter 1, p. 201.

[2] This experiment was carried out by Patrick Bestgen, Computer Engineer and Research Director, Consult'It Switzerland. He read the drafts of my thesis and volunteered to test my idea directly with ChatGPT.

Figure 3.3: Kanizsa illusion after performing a random chaotic modification. We still see the triangle, but ChatGPT does not.

> HUMAN: I don't see a six-pointed star. In any case, they aren't symmetrical! Don't you see another, more obvious shape?
>
> ChatGPT: You're right, if we're looking for perfect symmetry, the six-pointed star is not the most obvious shape. Looking at the white space, I more clearly see a triangle with rounded/incomplete tips.

It is certainly tempting to believe that, with sufficient training, ChatGPT will eventually provide adequate responses. The Kanizsa illusion, however, suggests that there is a significant difference in how data appears to the human mind and how it is processed by the machine. While the sets of objects are the same — namely, points on a screen — they would not be perceived through the same filters.

Artificial intelligence sees isolated points, each disconnected from the other; and it is quite a task to link them, to classify them, to identify in these disparate clouds the shape of a letter, a face, or a figure. At most, it can perform blind statistical associations, mimicking the expected responses. The human mind, on the other hand, perceives generalizable geometric patterns; it sometimes even distorts its perception to match the structure of its concepts, such as adding lines that do not exist. Moreover, it cannot help but see these geometric shapes emerge in the white space of the page, even with conscious effort. The illusion imposes itself with absolute necessity.

The position I am contemplating, that our *a priori* concepts alter our perception, lies at the heart of Kantian thought. It has often been described as a Copernican revolution[3]. It occupies an intermediate position between

[3]It is Kant himself who introduces this idea, though he does not mention Copernicus by name. What Kant has in mind, rather, is the method of Euclid in mathematics and that

idealism, which makes consciousness the measure of reality, and blind empiricism, which only acknowledges the reality of sensory experience[4]:

> *There are things given to us, as objects of our senses located outside us, but as to what they might be in themselves, we know nothing; we only know their phenomena, that is, the representations they produce in us by affecting our senses.*

Rather than naively thinking that we perceive the world as it is; rather than believing that the Universe is populated by squares, triangles, circles, matter, smells, or whatever else, Kant's position is to say that we only see what corresponds to the concepts within us. It is because we have, in our minds, the concept of a triangle that triangles appear to us in the tangible world. Instead of saying "Look, a triangle," perhaps we would do better to exclaim, like the surrealist painters: is this not a pipe? Does a triangle, in fact, exist anywhere in the Universe? What we take for one is, in the final analysis, just a collection of atoms suspended in the empty void of a white page.

The development of modern science seems to prove daily that our perceptions and concepts are far removed from physical reality. In Kant's time, the Universe was still that of weights and balances, atoms colliding, moved by Newton's forces in Euclidean space. But by the early twentieth century, our preconceived ideas about time, space, light, and gravitation burst open to such an extent that we must acknowledge that the world of today, the delicate edge of our thoughts, is no longer made of solid ground but of rocks, projections, and holes.

Not only is the Universe almost entirely made up of void, but the particles that populate it suffer from duality: their behavior, sometimes wave-like, sometimes particle-like, involves uncertainty in their trajectories (speed or position), so much so that we prefer to speak of probabilities of presence or quantum orbitals. Space-time itself becomes more complex: it bends under the presence of masses, contracts, and forms around monstrous black holes. Yet, we continue to see around us solid objects, endowed with colour and texture. It follows that our species must possess an intuitive representation of the world that is quite different from what physics teaches us to understand. Our sight, this eternal trickster, deceives us about the nature of things because we are simply incapable of seeing things as they truly are.

---

of Galileo in physics. He sees in the demonstrations of Euclid's Elements and in Galileo's inclined planes spaces where "reason only perceives what it produces itself according to its own plans" (Critique of Pure Reason, Preface to the Second Edition). He further adds: "It is in this attempt to change the method followed in metaphysics and thus, following the example of geometers and physicists, effect a complete revolution, that the work of our critique of speculative reason consists" (ibid.).

[4] Kant Emmanuel, *Prolegomena to Any Future Metaphysics*, §13, Remark II.

How do we see the world? And in what exactly do we differ from the machines we program? I have shown in previous chapters that there are three types of *a priori* structures in Kantian thought, which progress hand in hand: (1) intuitions, (2) the categories of the understanding, and (3) the Self.

Intuitions pertain to Sensibility. They are the pure forms of space and time, defined as the conditions without which it would be impossible to represent a phenomenon in time and space. These are prerequisites for any concrete experience of the world.

The twelve categories of the understanding, as their name indicates, belong to the Understanding. They allow us to isolate objects within the sensible phenomena, endowed with relations, and placed in relation to each other. Kant enumerates four general kinds of categories: quantity, quality, relations, and modality.

The Self, for its part, is the subjective foundation, the binding agent, the base of all perceptions. It is what makes our conscious experiences *our* experiences, and not simply isolated data in time and space. The Self is what gives our consciousness subjective unity: everything we feel, everything we experience, relates to us. This table, this book, this chair, the sensation of the wind on our cheek, the smells of the room, the sounds from the street — all of this forms a single whole.

If it is indeed the synthetic *a priori* judgement that is at issue when we seek to understand the difference between human reasoning and artificial reasoning, the following conclusion follows: AI continues to fail at ARC tasks because it does not possess the same *a priori* perceptual structures. It analyzes data, but these data do not manifest themselves in the same structures. What exactly could these structures be? How can we prove that the machine does not have them?

In the pages that follow, I will attempt to question, methodically, the extent to which a "machine" has — or does not have — the *a priori* structures we have identified. I will use the term "machine" to broadly refer to the Deep Learning tools used today, including neural networks. This common denomination — vague, I admit — should not lead the reader to forget that there is a great variety of algorithmic methods, each with its particularities. A review of the literature on theoretical computer science will be presented in chapters 7 and 6, where I will attempt to develop a rigorous answer to the questions I am only raising here.

Let us note, before concluding, that one could critique the notion of *a priori* judgement as presented by Kant. This definition indeed assumes the presence of innate structures, whose existence can be questioned. I will show in chapter 4 that the intuitions of space and time are just as problematic when considered in humans as they are in machines. The aim will then be to construct a new definition of judgement that, while taking into account Kant's critical contribution, is more consistent with the recent development

of science.

In short, what I undertake in this chapter is still only a preliminary critique of artificial intelligence; I deconstruct the machine piece by piece, like an engineer eager to demonstrate that the "feats" of our algorithms harbor some grey areas. However, I do not possess the theoretical tools necessary to understand these grey areas and theorize them. That is why, in the second part of this thesis, I will strive to rebuild the machine of the human mind piece by piece, retaining Kant's contributions but also considering recent scientific progress. Only this unified theory will allow us to understand in depth the differences between artificial intelligence and human intelligence.

3.1 Does the Machine Lack Intuition of Space?

It is often said that a machine does not think by itself; that it merely repeats the answers it has been carefully taught. It would not exhibit imagination, flexibility, or the ability to understand its own cognitive processes in a meta-reflective way. In this pessimistic scenario, it is the machine's very capacity for judgement that is called into question. One could indeed argue that algorithms are incapable of constructing the concept of an object by reducing the diversity of sensible sensation to a categorical unity. In Kant's language, it could be said that it lacks a henological function; or that the machine cannot subsume the data under a concept. This incapacity may, in turn, be due to at least one of these three things: a lack of sensation (no synthetic judgement), a lack of reason (no analytical judgement), or the absence of *a priori* forms.

The most radical option, and perhaps not the least plausible, is to say that the machine lacks sensation as such. The way it receives and processes data, though resembling the workings of the human mind (or at least the idea we have of it), is nothing but an efficient imitation. This would mean that the functioning of our sensory organs, in the final analysis, has nothing in common with the somewhat simplified and reductive processing of the computer equipped with sensors. This is an approach worth considering, as it calls into question a relatively unchallenged assumption in cognitive science (or at least not sufficiently), according to which perception is essentially an electrochemical processing of information comparable to the manipulation, at each moment, of sequences of «0» and «1». Gerald Edelman, Nobel laureate in Physiology, does not hesitate to remind his contemporaries of the philosophical naivety hidden behind such an assimilation[5]:

> *In many scientific circles, there remains a widespread belief that the brain is a computer. This belief is mistaken for a number of*

[5] Gerald Edelman, *Consciousness, Body and Brain.*

> *reasons. First, the computer works by using logic and arithmetic in very short intervals regulated by a clock. As we shall see, the brain does not operate by logical rules. To function, a computer must receive unambiguous input signals.*

There is one word in this passage that deserves, I believe, our full attention: ambiguity. The computer receives exact information. Even when a blurry image is presented to it, it transforms that image into a matrix of numbers. These numbers have no immediate spatial properties. The point located in the lower left corner is not *a priori* any farther from the lower right point. It is certainly possible to equip the machine with a metric and ask it to calculate a distance, but this metric is a choice made by the programmer. The only real, intrinsic differences between the points are given by numerical values: three numbers for the colour (red, green, blue), two corresponding to the position of the point on the photo (abscissa and ordinate), one number for the brightness, one for the albedo, and so on.

In contrast, for the human being, it could be that sensation operates through *a priori* forms of space and time. These forms, whose existence remains, of course, hypothetical, manifest themselves in that, even if the photo we are looking at is flat, we perceive perspectives. We distinguish the foreground from the background, we see depths appearing, we imagine trajectories in three-dimensional space, and so on. For example, if a shape is presented in the background, we unconsciously enlarge it in our perception to account for the perspective effect. This means that our information is not exact, but is influenced by the context.

This unconscious process is particularly evident in the Ponzo illusion (see figure 3.4). In this illusion, you are asked to compare two horizontal bars and determine—which bar is longer. They are the same size, but the lines of convergence in the background create the illusion of perspective, so that the upper bar appears longer. It is as if the human mind, unwittingly, projects what it sees into a three-dimensional perspective. The interesting point is that the illusion occurs even when one knows that the bars are the same size.

Within algorithms, it is not certain that an *a priori* form of space exists. Faced with the Ponzo illusion, the program will merely use its metric to measure the exact length of the two line segments. It will never be wrong. It could also measure what the size difference would be if these lines were indeed part of a three-dimensional space. For this, it would simply need to use the laws of perspective. For it, it would be entirely indifferent to use the first approach or the second. It does not have a privileged view of the images it analyzes.

It is certainly possible to train the computer to respond like a human to the question: "Which bar is bigger, according to you?". This can be done by giving it a large sample of images, accompanied by the answers

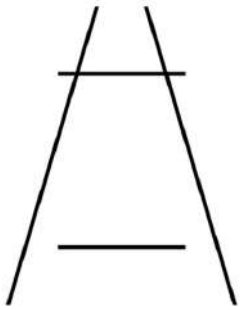


Figure 3.4: Ponzo Illusion: The two horizontal lines are the same size, but the one at the top of the image seems larger.

from many humans, and asking it to build a statistical model associating the correct answers with the images viewed. However, it does not seem that this illusion, to which it will claim to be subjected, is experienced as an intrinsic necessity, testifying to a particular way of structuring its perception of the world. It behaves like the parrot that has been taught to say it likes reading Shakespeare.

What is extraordinary about the Ponzo illusion, and which seems to place human perception in a different register from the statistical analysis of the points in the image, lies in two factors: (1) the prevalence of the phenomenon and (2) its inevitability. By *prevalence*, I mean that the perspective effect appears even in the face of abstract images, obtained by the nearly random arrangement of straight segments on a blank page. It is true that these lines must be placed in such a way as to respect the laws of perspective, but our gaze is so skilled at detecting perspectives that the designer will not need much effort. It is as if the human being absolutely seeks to place everything that falls before its eyes into a spatial intuition. It cannot help but do so.

This brings us back to point (2), which is *the inevitability* of the phenomenon. Even after realizing its mistake in broad strokes, it remains rigorously impossible for the human being to avoid being a victim of the optical illusion. It can, in a final effort to rid itself of the problem, turn the image 180 degrees, in which case the illusion indeed diminishes significantly. But as soon as it is placed back in its original position, it reappears as surely as a stone, dropped from the top of a precipice, will fall to the Earth. If the illusion were linked to a simple mistake in rational judgement, similar to a miscalculation, it would suffice to be aware of it to correct it. But here, everything seems to happen as if the root of the problem is prior to any possible perception.

We thus come to an important methodological remark: to determine the existence of *a priori* forms in the thinking of a human or a machine, it may be more instructive to focus, not on what they are capable of doing, but

on what they are fundamentally incapable of doing. I call this principle the *subtractive principle of intelligibility.*

Définition 1 (Subtractive Principle of Intelligibility)**.** *It is by studying what the mind is fundamentally incapable of, through no conscious effort, and not what it is capable of, that one can identify the presence of an a priori structure.*

Based on this principle, I would now like to ask whether, analogously to what I said for space, there is a deficit in human perception that would testify to the existence of an *a priori* form of time? In Kantian theory, in fact, the pure intuition of space is accompanied by a pure intuition of time, which concerns not external sense, but the phenomena internal to the subjects. It is also time, as we saw in section 2.4.4 on schematism, that is at the origin of the judgement allowing the unification of various sensations under a single conceptual unit. It thus plays, so to speak, a more fundamental role than space. Its analysis, moreover, proves to be more complex.

3.2 Does the Machine Lack Intuition of Time?

I showed in the previous chapter that the machine does not share the same *a priori* structure of space, but I left in the shadows the question of how it stands with respect to the conception of time. As a first observation, let us note that if human perception of time were perfectly objective, events would unfold with the same rigor as a metronome. We would be able to count the seconds between two flashes of light, two sounds, two movements, without ever erring; the phenomena of consciousness would succeed one another in a linear and homogeneous flow. Perhaps this is indeed the case for certain fictional individuals, such as Borges' Funes, who is capable of giving the exact time at any moment. Whether running, speaking, or jumping, he behaves like a walking clock[6]. Kant himself came quite close to this caricature, if we are to believe the writer Thomas de Quincey[7], who reports that the neighbors in his district would set their watches by his daily walks.

And yet, in reality, certain perceptual illusions challenge this conception. The most well-known is arguably chronostasis, also referred to as the stopped-clock illusion. This illusion manifests as the sensation that time stretches out, that it lasts longer than it should. One experiences it when, after listening to the regular "tick-tock" of the clock while gazing out the window at the landscape, one suddenly turns their eyes to the face of the clock. In that instant, it may feel as though the second hand imperceptibly freezes, lingering just a bit too long in the position it held when caught by

[6] Borges, *Funes, His Memory*

[7] Thomas de Quincey, *The Last Days of Immanuel Kant*, Mille et une nuits, 1996, pp. 27–33.

the gaze. After this brief dead time—uncannily reminiscent of a glitch in the intricate matrix of the universe—the hand seems to resume its implacable motion.

I would like to examine this perceptual illusion in order to determine whether the human mind differs, or not, from the behavior of a digital machine. To pursue this endeavour, I will begin by reviewing the neuroscientific explanations commonly offered for chronostasis. In the mid-twentieth century, when researchers first became interested in the phenomenon, they often—though implicitly—assumed that the brain functioned like an information-processing machine. This machine receives images from the external world, modifies them at will, before projecting them onto the "theatre of consciousness." Yet in the early twenty-first century, a number of experiments began to cast doubt on this computationalist paradigm, such that the mechanism of chronostasis remains, even today, largely unresolved. We have clues, to be sure, but no consensus.

In this chapter, I will suggest that by adopting a Kantian perspective on the phenomenon—and by relinquishing the idea that the brain behaves like a computer—one may gain a compelling insight into the problem. The analysis I will develop aims to defend the idea that time is not perceived objectively, as it would be when measured by a clock, but rather appears in our consciousness as a whole in which present, past, and imminent future blend into one another. It is only through a kind of partitioning that this amalgamated temporal experience is spread out across a timeline endowed with orientation. I will emphasize the notion that this partitioning is an *a priori*—a structure of the human mind. It functions as a kind of filter. When phenomena pass through this filter, they take on the appearance of temporal objectivity; but there are specific cases in which this construction dissipates like an illusion.

3.2.1 The Saccade Model

Chronostasis has long fascinated scientific observers, sparking a variety of hypotheses in the minds of even the most distinguished researchers. Until the late 1990s, the dominant explanation attributed this dilation of subjective time to the rapid eye movements known as saccades. It is indeed well established that, when the eyes move briefly, your vision becomes blurred for a fraction of a second. To compensate for this effect, the brain corrects the perception by substituting, in place of the blurry images you actually see, clear images available either before or *after* the saccade. You read that correctly: the brain replaces the blurry image seen at instant t with a clear image seen at the instant just *after*. It is thus a visuo-spatial correction performed *a posteriori*.

Let me now specify the exact functioning of this neuro-ocular mechanism. Before diving into technical details, I would like to note that the saccadic

phenomenon I'm speaking of is something you experience every day, even when you are at rest. You are likely unaware of the micro-movements of your eyes happening even now, as you read these pages—and yet, as the saying goes—and yet it moves. Far from remaining quietly still, your eye is constantly oscillating within its socket. What appears to be a smooth and uninterrupted perception of the world—a calm, cinematic river of vision—is in truth a chaotic, capricious, and uncertain torrent, littered with countless micro-corrections.

The epistemological framework underlying this saccadic mechanism is akin to that of a movie theatre. Your brain is a learned machine that receives images, performs calculations, and makes adjustments to those images in order to present to its master—consciousness—a view of the world that is as clean, as aesthetic, and as practical as possible. When it does its job well, the images are placed on a long reel and played one after another. Consciousness, comfortably seated, watches the final result, unaware of the backstage constructions that made it possible. The film of consciousness plays on endlessly, always at the same speed. An image imprinted at the beginning of the reel will be perceived as having occurred before an image imprinted at the end of that same reel.

This model I'm describing—let us note right away—is not exactly the one employed by neuroscientists. Nevertheless, I believe it captures the underlying epistemological intuitions behind neuroscientific explanations of chronostasis. During a saccade, it is as if the brain receives a batch of damaged images. The question then arises: what should it do with them? If it engraves them onto the reel in the order in which they were received, consciousness risks becoming disoriented. That's why the brain prefers to discard the blurry images and duplicate the ones that are still clear.

To summarize the situation, there are four events that occur on the reel of consciousness during a saccade: (1) the moment you begin the saccade, (2) the moment your eyes are in motion, (3) the moment you fixate on the clock, and (4) the moment you see the hand move. It so happens that between events (2) and (3), your vision is blurred due to ocular motion, so that the reel presents patches worthy of abstract art[8]. According to the saccade model, the brain must therefore apply a correction. It will replace, *a posteriori*, the blurry section of the reel with a clear image. Yet the only clear image it has is the one you perceive at instant (3). You are thus left with the impression that the image from instant (3) began well before (3), shortly after (2). This gives you the retrospective sensation that the clock hand remained still for longer than it should have.

[8]The reader will likely not notice this blurring effect when moving their head. Neurologically, the blur is explained by a decrease in visual sensitivity during the saccade. This effect is notably studied by Knoll, Jonas, *Spatio-temporal representations during eye movements and their neuronal correlates*, 2012.

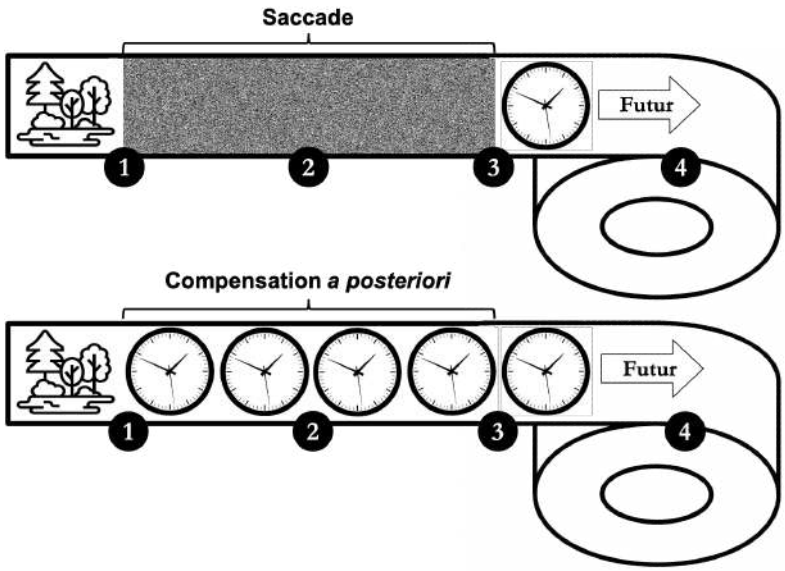


Figure 3.5: In the physiological model of chronostasis, time is an objective structure comparable to a reel that records our sensory perceptions. When a saccadic movement occurs, the brain replaces the resulting blurred image with the first clear image observed at the end of that saccade. This compensatory effect is *a posteriori*.

3.2.2 Objection: Auditory Chronostasis

The explanation of chronostasis that I have outlined above rests on a foundational assumption: that chronostasis always occurs in tandem with a saccade. Yet everyday experience seems to contradict this presupposition. Chronostasis can arise in non-visual contexts—for instance, during a phone conversation. If you shift the receiver from one ear to the other in order to reach for your calendar, you may have the impression, upon picking up the thread of the conversation again, that the interval corresponding to this brief interruption lasted far longer than expected. It gives you the uncanny sense that your interlocutor, somehow intuitively sensing that you couldn't listen attentively at that precise moment, paused to wait for you. You distinctly feel that they went silent, or that the line was momentarily cut.

This auditory form of chronostasis received experimental support in the early 2000s. In a study conducted at the University of Oxford, a group of subjects was asked to wear headphones and assess the duration between two auditory stimuli[9]. Participants were instructed to press a key as soon as a beep began and again as soon as it ended. The goal was to measure the gap between the objective time of the beep (recorded by the computer) and

[9] Iona Hodinott-Hill, Kai V Thilo, Alan Cowey, Vincent Walsh, *Auditory Chronostasis: Hanging on the Telephone*, October 15, 2002, pp. 1779–1781.

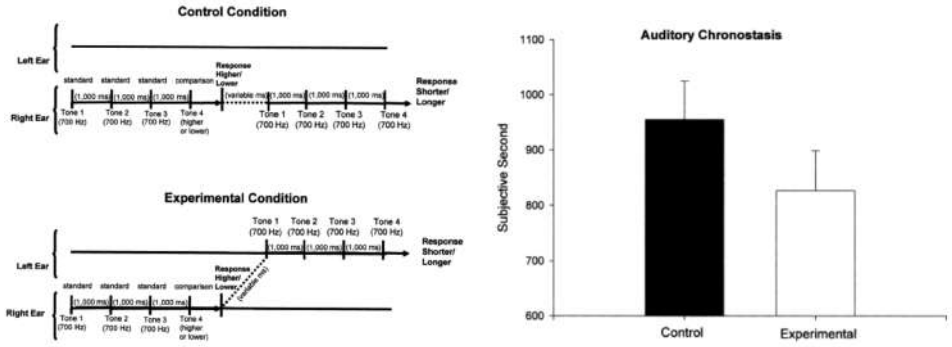


Figure 3.6: In this experiment conducted at Oxford, subjects were asked to determine the duration between two auditory beeps. In the control condition, all stimuli were delivered to the right ear (top left image). In the experimental condition, the first four beeps were delivered to the right ear and the next four to the left ear (bottom left image). This shift in attention altered the perception of time (image on the right).

the subjective duration perceived by the individual, in order to identify the precise conditions under which this discrepancy emerges.

In the control condition, four consecutive beeps were delivered to the right ear, each separated by one second. This first sequence was followed by a variable waiting period, after which four more beeps were played—again, one sound per second. Participants were then asked whether the time interval separating the first four beeps was the same as the interval between the latter four. To avoid any confusion, the last of the initial four beeps was deliberately played at a lower pitch, in order to signal to participants that the following beeps belonged to the second series. In this control condition, subjects did not experience any distortion in temporal perception. Their evaluations of the intervals separating each of the four beeps in both sound sequences consistently hovered around one second.

In the test condition, the setup was modified to induce the illusion of chronostasis. The first four beeps were played in the *right* ear at one-second intervals, and the next four in the *left* ear (still at the same frequency). This new protocol was designed to model, in a laboratory setting, real-life situations in which an individual switches the phone from one ear to the other. Participants were informed in advance that the sound would switch ears, yet they still fell prey to the temporal illusion: they perceived the intervals between the beeps in the second sequence as being stretched.

This result is consistent with the expected effect in cases of chronostasis. The Oxford experiment thus suggests that the optical illusion, contrary to what was previously believed, is not explained—at least not exclusively—by

the physiological mechanism of saccades, but rather belongs to a broader phenomenon. More fundamentally, it seems necessary to reconsider the model positing an internal clock within the organism. Time does not appear as a reel awaiting the imprint of our various perceptions. How, then, can we explain the perception of time? Is there an interpretation of chronostasis that is compatible with the Kantian idea that there exists, in the human mind, an *a priori* intuition of time?

3.2.3 Hypothesis: Differential Perception

Since the Oxford experiment undermined the explanation based on the mechanism of saccades, scientists have sought alternative models. Several avenues of inquiry have been proposed in the literature, which cluster around three main methodological approaches:

- **Motor Activity:** Some authors continue to maintain that, even if chronostasis manifests outside of ocular saccades, the phenomenon remains linked to a movement of the organism. In the Oxford experiment I mentioned earlier, subjects do indeed move their index finger to click a button.

- **Attention:** Some researchers argue that the cause of chronostasis might instead be related to attentional shifts. It is not the movement of an organ that proves decisive, but rather the fact that our concentration moves from one part of the room to another, or from one ear to the other[10].

- **Change:** Since the 1950s, studies have shown that time perception is influenced by the nature of the stimulus itself[11]. It has been especially noted that contrasts within the stimuli seem to accompany the emergence of temporal illusion.

While each of these scientific hypotheses offers a promising path toward clarifying the precise physiological mechanisms behind chronostasis, they do

[10] In one experiment, Rose, D., and Summers, J. generate sequences of flashing lights and observe the presence of chronostasis when the first flash appears and, in some cases, when it disappears. Subjects do not move their heads or eyes, which suggests that saccades are not responsible for the phenomenon. See *Duration illusions in a train of visual stimuli*, Scopus, Volume 24, Issue 10, 1995, Pages 1177–1187.

[11] See in particular 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. See also Schiffman H.R., Bobko D. J., *Effects of stimulus complexity on the perception of brief temporal intervals*, J Exp Psychol, 1974.

not necessarily unveil an *a priori* structure of time perception. These hypotheses merely establish *a posteriori* correlations between various empirical variables. The entire experimental methodology consists in observing how the alteration of a factor such as attention induces an increase or decrease in the perceptual illusion—and in some cases, suppresses it altogether. Yet according to the principle of subtractive intelligibility I defended in the previous chapter, the existence of an *a priori* form of perception can be revealed only by the presence of an illusion that resists all *conscious effort*.

The method to be adopted, in the context of my inquiry, thus consists in determining, among the four scientific hypotheses that have been raised, which are those that cannot be controlled by a *conscious effort*? It seems possible, with sufficient willpower, to control one's attention, or to choose whether or not to move one's head, eyes, or hands; it is, on the other hand, strictly impossible to control the change that occurs between two perceptions. No matter how one turns the problem, when your gaze shifts from contemplating the landscape to observing the needle on a dial, or when your phone moves from your left ear to your right, there is *de facto* a change in the content of your perception—over which you have no control.

Since my objective is to determine whether there exists an *a priori* structure of time, and under what conditions it manifests, I would like to pursue further the hypothesis that links chronostasis to change, so as to give it a philosophical interpretation. As we saw in the case of space, the human mind does not seem to apprehend spatial points in isolation; these points possess reality for us only through the relations that connect them to other points in space. In the same way, I would like to suggest that instants do not exist individually in human perception. All our mind can truly observe is the change that occurs between two events.

To shed light on this idea with a clarifying analogy, we may think of the notions of speed and position in the context of physical trajectories. When a body moves through space, the naive conception is to say that it occupies, at each moment, a specific position with a specific speed. But what if instants did not exist? Even for a macroscopic object like a ball, it is not straightforward to measure with precision its position or speed at a given moment. If you take a picture of its trajectory, the ball will appear blurry, requiring you to adjust your camera's shutter speed. With the right settings, you may hope to capture the ball's exact position, but on that very image, there will be no trace of its speed.

Indeed, the speed of a moving object is not an intrinsic property of that object, but a differential one, defined as the ratio between the distance traveled and the time taken to travel that distance. If the ball takes one second to go from point A to point B, and these points are twenty meters apart, it travels at twenty meters per second. This is not instantaneous speed, but average speed over the whole trajectory.

To know the speed at a more precise point in the movement, you need

to bring points A and B closer together, measure how long the ball takes to travel not twenty meters, but twenty centimeters, then twenty millimeters, twenty nanometers. As you continually shrink the distance between A and B, you approach an increasingly precise approximation of speed, which we call instantaneous velocity. This speed, however, is a mathematical abstraction. There is no guarantee that such a magnitude exists beyond the fiction of our theories. Heisenberg's principle even seems to suggest that, at quantum scales, this ideal of instantaneous speed or position collapses.

What I wish to illustrate with this example is that the perceptual contents of consciousness are not necessarily absolute magnitudes, existing at each instant, but depend on the other events occurring in the immediate temporal vicinity. This notion of "vicinity" is deliberately kept vague for the moment. What I want to express, in essence, is that the human being does not apprehend events sequentially, ushering instants into consciousness like so many people lining up at a door. The great river that is time is not made of indivisible droplets, like billiard balls. It is not the strip of a great photographic film meant to illuminate the "theatre of consciousness." On the contrary, subjective time presents itself as a whole that our mind transforms into a chronological sequence with a beginning and an end.

Could this homogeneity, this totality in which time appears to us, be at the root of the vertigo, the stupor, the confession of helplessness that grips us when we try to apprehend the present moment? Such a moment seems not to exist—or at least resists even the most attentive philosophical scrutiny. Before we think it, it is a future; and the moment we think it, it becomes a past—so that the instant, the past, and the immediate future are entangled with one another. They reflect one another in an eternal hall of mirrors.

Instead of conceiving time as an infinite ribbon that the mind scans linearly, I would like to propose another geometric representation of time. Strictly speaking, this is merely an analogy; but I believe it may offer insight. Since we do not perceive instants in themselves, but rather the changes occurring in a temporal neighborhood, this neighborhood could be likened to a small-radius sphere. Our perception of these amalgamated phenomena would then consist in projecting this sphere onto a temporal line running from past to future. Depending on the orientation you give to the sphere before projecting it, it may happen that two events A and B occupy the same position on the timeline. Simultaneity, therefore, is not an absolute (see figure 3.7).

This phenomenon I describe is well known to sound engineers. A major issue they face during projection or editing is synchronizing audio with image. Fortunately for them, a delay of up to one hundred milliseconds will not be noticed by a viewer. The reason is that visual and auditory information are not processed at the same speed by the brain. Even when we observe people walking down the street, there is always a lag. The brain therefore

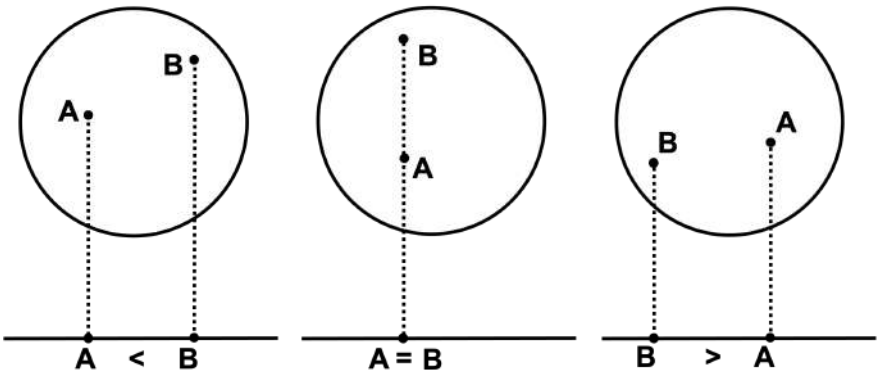


Figure 3.7: When projecting two points A and B from a sphere onto a line, it is possible to orient the sphere in such a way that (i) A precedes B, (ii) A and B coincide, or (iii) B precedes A.

constructs the temporal structure of events *a posteriori*, based on different causal signals. The sound of a fall must be caused by the object hitting the ground. A correction is made.

A scientific validation of this idea can be found in a study conducted by University College London, in which participants were asked to press a button to turn on a light[12]. The setup was designed so that there was a 100-millisecond delay between pressing the button and the light turning on. At first, participants sensed the delay, but after some time adapting to the system, they ended up perceiving the two events as simultaneous. If, at that point, the button and the light are truly synchronized, subjects then report the strange impression that the light appears *before* they press the button—in other words, there is a reversal of action and sensation. The authors of the study note that similar desynchronization effects also occur when the participant is passive and does not act by pressing a button, but in that case, the intensity of the phenomenon is reduced by 50

3.2.4 Subjective time vs Objective time

The phenomenon of chronostasis, as I have analysed it, thus supports—or at least remains compatible with—the interpretation that there exists a perceptual structure of time in the human being, one that is not comparable to the processing of an automatic machine. The studies I have discussed seem, in particular, to confirm the idea that simultaneity and succession of events are not absolute realities, but constructions. I must nonetheless

[12] Haggard P., Clark S., Kalogeras J., *Voluntary action and conscious awareness*, Nature Neurosciences, 2002.

emphasize an important point: this perceptual construction that I have in view is not carried out *a posteriori*, but functions rather as an *a priori* filter.

What I mean by this is that the brain does not simply take incoming events—such as a sound or an image—and pin them one by one along its timeline. To say so would be to believe that time is an objective feature of the world, a line drawn in indelible ink across the Universe, which our brain merely attempts to reconstruct. This linear vision of time, inherited from Newton, is a naive one. It does not take into account the recent advances of physics[13]. Moreover, the great lesson of Kant, as I have expressed repeatedly, lies in insisting that it is objects that conform to our structures of time and space—not the other way around. One must not suppose, then, that time exists as a fundamental reality of the world, but rather admit that both objective and subjective time are structures of our thought.

Chronostasis is therefore not—as is often claimed—a distortion of the objective reality of time, attributable to our subjectivity. It is rather the sign that neither objective time nor subjective time adequately represents the events of the world. What exactly do I mean by this? When you attempt to draw three-dimensional objects, such as the Earth, on a flat sheet of paper, distortions inevitably occur. Thus, the Mercator projection alters the volume of continents while preserving angles, whereas the Peters projection sacrifices angular accuracy in order to preserve surface areas. In the same way, objective time and subjective time are two ways of projecting, into an intelligible plane, the events of the world.

Objective time is obtained by taking a regular movement as a frame of reference. This may be, for instance, the movement of a needle across the finely-oiled dial of a clock, or the radiation emitted by the cesium-133 atom. Yet it is by no means certain that these phenomena are perfectly regular in any absolute sense. They are, in the end, merely hypotheses—effective and convenient, to be sure—but unverifiable. If the entire world were to slow down, if all movements occurred more slowly, would we be able to perceive it?

Subjective time, on the other hand, is constructed by the mind through measuring the degree of change or the discrepancy between two perceptions. When your eyes shift from gazing at a pastoral landscape to observing the needle of your watch, you experience more changes than when you watch the needle quietly move from one position to another. There is, indeed, a sharper contrast between sunlit wheat fields and your metallic watch face than when your eyes remain fixed on the hands themselves. Subjective time preserves, so to speak—if I may extend this cartographic analogy—the volume of events, whereas objective time preserves their distances.

It is important to note that it is the entire set of mental contents that

[13]For a critical and highly accessible review of the notion of time in physics, see Peter W. Evan, *A Study of Time in Modern Physics*.

must be taken into account in this temporal measurement, and not only the visual field. It may be that you are staring at the needle, but your mind has drifted so far afield that time seems endless. When you are waiting for the bus, time appears slow, because you are constantly expecting to see the bus emerge around the corner. You look in the direction from which it should come, you imagine it, it is there in your mind's eye, ready to greet you—then you realize the street remains irredeemably empty. Not a second has passed before you imagine it again. This rapid contrast between the idea you form (your desire) and the image of reality generates such a tension that the lived time far exceeds the mere movement of the needle.

The fact that objective and subjective time are two possible projections of a single amalgamated reality does not mean that they are of equal value. Just as the Mercator projection has prevailed in most world maps—due to its maritime efficiency—the objective definition of time asserts itself through its greater precision of measurement. Yet it is not inconceivable that the subjective representation of time holds interest in certain contexts. There may, for instance, be an evolutionary advantage to our species not being endowed with an objective vision of time: the temporal slowdown we experience in certain emotional situations—when faced with imminent danger, for example—may be linked to our ability to process information and its multiple connexions as a hyperconnected whole (rather than as a mere sum of individual stimuli). The aim of this thesis is precisely to understand the limitations encountered by algorithms, which sequentialize time and space, when confronted with problems that demand a more abstract, more holistic, more general analysis—such as the now-famous ARC tasks.

To summarize what has been said, one might suppose that the human mind—regardless of the information it encounters—can only apprehend this information within a pre-existing spatial and temporal structure. It does not perceive the points of space and time in isolation, like so many blotches of colour thrown flat across a ribbon of paper. On the contrary, it conceives them holistically. The points acquire intelligibility through the arrangement they occupy in relation to one another; that is to say, through the difference that separates them. This difference may be understood in a spatial sense (a distance between two points), or in a temporal sense (a contrast between two events).

In contrast to the human mind, the computer takes isolated points as raw data, independently of the relations that link them together. It is not subject to optical illusions, yet it proves incapable of effectively connecting these points in a way that grants them meaning. Its functioning—purely sequential—makes it less apt at grasping the whole. One might say that it operates according to an atomic structure of time and space: phenomena are sequenced and decomposed so as to appear as a conjunction of disjointed bits of information.

Although these observations constitute entirely reasonable critiques, it

may be excessive to claim that the machine, merely because it does not process information according to the same spatio-temporal structures as the human being, is incapable of judgement. It is plausible to think that there is indeed such a thing as judgement within the machine, understood as a capacity to manipulate concepts—but these concepts are purely rational. In the absence of the pure forms of sensation, which weave intrinsic relations between objects, the algorithm must be content to link ideas to one another through logic alone. In the language of Kant, this amounts to saying that the machine is a subject for which objective deduction fails. By "objective deduction," let us recall, Kant seeks to understand how it is that the categories of the Understanding can apply to something, rather than to nothing. In the absence of sensation, these logical categories do indeed exist, but they spin in a vacuum. It is a mill whose blades turn without ever having any grain to grind!

3.3 Does the Machine Lack Contemplative Reason?

I have suggested that the machine does not process spatial and temporal information according to the same *a priori* structures as the human being. If this line of reasoning is to be believed, it would imply that the machine lacks a properly defined faculty of sensation. In Kant's philosophy, sensation is indeed shaped as the act of representing phenomena within a pre-established space of objectivity. When we see the colour red, we do not see photons and waves as they are *in themselves*. Do photons even have any known form? How could we ever access the true nature of things, as they supposedly exist beyond the representation we form of them in our minds? Physical bodies never appear to us except through forms that depend on our prior perceptual structures. It is because human intuition is spatial and temporal that all physical phenomena are presented to us in time and in space. Such has been the case since humanity began. And it will remain so—even after one were to prove that space and time possess no physical reality outside of our subjective sensation.

From this analysis, it follows that the machine does not represent phenomena in a subjective form, but sees them solely through the lens of objective linguistic terms: numbers, symbols, definitions determined in advance. One could say, figuratively, that reality is flattened. What is lacking is an ontological depth, a differentiation between the signs used to speak of objects and the idea of the objects themselves. As a consequence, algorithms do not relate to phenomena but to numerical symbols. They process information syntactically, as a set of objective and manipulable numbers, without possessing any semantics.

Of course, the question of what an artificial being would feel—if it ever

came to declare that it experienced emotions—would require one to step *into its head*, and is metaphysically undecidable. I am not here attempting to say anything about the machine's subjectivity, but merely to emphasize that it lacks a certain structure in its perception of information—one that could, functionally speaking, be assimilated to what we know of human sensation.

Having discussed the problem of sensation, the next question I must examine is whether the computations specific to the machine fall within a rationality in the full sense of the term—or whether they belong to something else entirely. In other words, I would like to ask to what extent we can think of algorithms as being endowed with reason. The position which claims that the machine lacks reason, that it merely imitates logical thought—more or less faithfully—seems rather surprising. According to common sense, it is generally more readily admitted that the computer lacks sensation than abstract reflection. This is due to the fact that the first machines were calculating machines, and calculation is often seen as the very archetype of rationality.

And yet, if we adhere to Kantian theory, reason is not defined *solely* by the manipulation of symbols, but *primarily* by the capacity to ascend toward the unconditional principles of truth. In this respect, Kant is indebted to the thought of Plato, who saw in *noèsis* this very capacity to seek the anhypothetical: beginning from simple hypotheses and rising step by step, the intellect seeks to grasp the primordial unity of the world—to understand what stands behind, or beyond, all principles. This unity, in Plato's *Republic*, is the Idea of the Good, which is often compared to the sun. This idea is that without which no other idea could be conceived. It is like the light that allows the eyes to see the world. That is, in essence, what is said in this passage from Book VI of the *Republic*:

> *Who then, of all the gods in heaven, is it whose light allows our eyes to see in the best possible way? [...] It is the sun, obviously, that you are asking me to name. [...] Just as it is right to think that light and sight resemble the sun, but wrong to think they are the sun, so too, in the intelligible world, it is right to think that knowledge and truth both resemble the good, but wrong to think that either one is the good*[14].

To say of the machine that it lacks Reason is not, of course, to deny it the purely factual ability to produce deductions from symbols. It is self-evident that algorithms, by their very structure, are capable of deriving logical equations. What I would like to challenge here concerns only a particular dimension of reason—one that distinguishes it from a mere rationality. For reason is not a faculty of calculation, but a certain aptitude for the contemplation of ideas. In intellectual life, there is an experience of clarity, of lucidity,

[14] Plato, *The Republic*, Book VI.

which consists in seeing one's own processes of thought as if they were an object in the world. This experience is intimately bound to an aesthetic feeling.

Scientists who have passed through such states report having felt something akin to a kind of intellectual grace. In the interstices of an equation, it is a fragment of the Universe that is suddenly revealed to them. The mystery of the world, the immensity of the whole, the structure, the order, the intrinsic harmony of geometric forms, the daze brought on by an idea whose content seems to exceed what the human mind is capable of pouring into it—this is what gives rise to a metaphysical vertigo. The essence of the world seems to unveil itself, for a few moments, to the one who, by the grace of prolonged study, sets foot within the clarity of concepts.

This visual—almost existential—experience is what one might properly call intellectual intuition. Just as spatial and temporal intuition give form to perceptual phenomena by predetermining the way an event appears to us, intellectual intuition would give abstract ideas an essential form, allowing us to grasp by intellect that which has no hold in the sensible world. If the position I am evoking here may seem mystical at first glance, one need only think of the idea of the infinite. There are concepts present in the human mind whose essence exceeds the understanding, yet of which we may still form a clear and distinct idea. When you think of the set of natural numbers, you would be hard-pressed to visualize them all at once. If I tell you that the number $2^{136.279.841} - 1$ is the largest known prime number, you may well feel dizzy upon learning that it contains $41,024,320$ digits. That dizziness one feels when trying to represent such seemingly mundane expressions as "the set of prime numbers" brings us asymptotically closer to a more essential, more rational, more intellectual understanding of an infinity that cannot be represented through the senses—an understanding that borders on aesthetic contemplation.

It is worth noting that the idea of intellectual intuition I examine here is not grounded in Kantian theory, but in Platonism. For Kant, indeed, intellectual intuition does not exist. On this point he could not be clearer. His *Critique of Pure Reason* is a plea against the notion that ideas have an existence independent of sensation. Pure reason, according to him, may certainly speak about moral things, but it cannot lead to scientific knowledge in the strict sense. Kant rejects Plato's *noèsis*, conceived as the search for the anhypothetical.

From Kant's point of view, then, there exists in any case no speculative intellectual intuition in human judgement. The concepts of God, the Universe, the infinite all collapse into contradictions. To simplify somewhat, one might say that these concepts are symbols similar to those manipulated by a soulless machine. It is certainly possible to establish rules regarding the use of these symbols, just as one might use rules to translate an unknown language into a known one. But this does not grant the symbols content—a

concrete semantics. In the final analysis, the only content of scientific and philosophical theories is given through sensible experience. This is Kant's key point. Which means that, from this second perspective, the machine would possess no less speculative reason than the human being.

I leave it to the reader to decide whose ideas they prefer to adopt—those of Kant or of Plato. In the Kantian view, contemplative reason exists only as pure moral and aesthetic sensation. In the Platonic view, it exists as a full-fledged intellectual experience. In either case, we are offered an interesting criterion to distinguish the human mind from the functioning of the machine, one that is closely tied to the intuition of infinity. I will return to this idea later.

3.4 Does the Machine Lack Understanding?

In light of what I have argued thus far, it seems reasonable to say that the machine possesses a form of rationality, but lacks sensitive structures. It can indeed manipulate symbols in a logical and deductive manner, and produce conclusions within the realm of pure speculation, but it does not possess the intuitions of space and time. Does this mean, then, that the machine feels nothing at all?

I would like to soften that hypothesis by defending the idea that it possesses just enough sensibility to produce synthetic *a posteriori* judgements, and just enough reason to carry out analytic judgement, but not synthetic *a priori* judgements. In other words, I would argue that the machine does indeed possess both sensibility and rationality, but not understanding — understood here as the faculty for judging truth in the absence of extensive experience or *a priori* logical justification. This nuanced stance I wish to advocate seems to me advantageous in several respects.

First of all, it offers a more charitable view of artificial intelligence. When it comes to critiquing a scientific paradigm, it is often wise to grant one's opponent a certain number of premises, and to explore whether, in spite of — or indeed thanks to — those very premises, one can still expose certain shortcomings. In the worst-case scenario, these shortcomings appear as contradictions between the premises themselves; but more often, it is simply a matter of showing that the initial project is not as promising as once believed. This may occur because it becomes necessary to introduce new premises, some of which are unconvincing, or because certain blind spots emerge in the thinking — shadowy areas, naïvetés in the way we imagine the human mind, its functioning, its essential determinants.

I want to design my thesis so that it invites debate, provokes controversy. When we deny the machine any capacity for sensation or reason, when we reduce algorithms to soulless automatons — akin to those described by Descartes, performing nothing more than simulations of human reasoning

— the discussion becomes futile, amounting to little more than pounding the table and dogmatically asserting one's convictions.

It is, after all, impossible to step into the shoes of a machine to determine whether it "feels" or "thinks." There is something here that is intrinsically unverifiable — and this, for reasons rooted in definition itself. Subjectivity, at least within our Western metaphysical tradition, is conceived as a strictly private and individual domain. Everything that occurs within my mind belongs solely to me, and to no one else. As a result, this mental space remains inaccessible.

I say this is a fact of our materialist metaphysics, and not merely a fact *in itself*, for there are cultures in which belief in telepathy or communication with the spirits of the dead is commonplace. Consequently, for them, the problem of subjectivity may not provoke exactly the same reflections, the same promises, or the same dead ends as it does for us[15]. Even in these cultures, to be sure, there remains an irreducible problem of communication. I merely wish to point out that it is perhaps not understood in quite so radical a fashion.

Let us note carefully: the problem of subjectivity arises just as much in relation to a machine as it does in relation to another human being. I cannot know for certain whether the reader whose eyes now fall upon these lines is truly feeling or thinking something — or whether they are merely trying to convince me so, with a performance that would make even the most seasoned actor pale in comparison. I therefore assume that my interlocutor is, like myself, capable of these faculties; and it is from this starting point that a properly philosophical debate can begin.

It may happen, in the course of conversation, that one interlocutor notices in the other a lack of sensitivity, logic, or life experience; and despite the most conscientious explanations, it may prove difficult to overcome the ineffability — the incommunicability — that haunts all human relationships. Whenever this occurs, the real question is: what is the element that prevents mutual understanding? Very often, it is a value. Human discourse is generally conditioned by underlying moral representations, concerning either the

[15]One might think in particular of shamanic cultures, in which dreams provide access to a subjective mental space shared by multiple intelligent beings — animals, plants, and the deceased alike — where communication is made possible. "All oral-tradition societies, especially those with shamanic practices, distinguish between two worlds: this one, of the everyday, the profane, the ordinary, and the 'other-world,' inhabited by gods, ancestors, ghosts, fantastical creatures, and beings from this world temporarily endowed with special properties, such as the shaman. The other-world is designated by each society with a term that may be translated as 'sacred.' Communication between these two worlds is deemed possible, and many misfortunes — sometimes fortunes — are considered to be the result. Beings from the other-world threaten or attack souls, bodies, possessions, or environments. But the other-world also offers humans the possibility of communicating with it and responding to its intrusions. This communication most often occurs through dreams, considered to be a voice emitted and governed by the other-world, thereby revealing the truth to come, delivering prognoses and oracles[16].

self or the world. Certain ideas remain opaque to us because of what they imply about our own self-definition. Some will reject this or that statement regarding the matter of men and women because they do not wish to appear, in their own eyes, as reactionaries; others will refuse to hear criticism of a political party or institution they have always supported, because such criticism — unknowingly — would render meaningless the time they spent defending that organization.

In some cases, the obstacle may also arise from a lack of aptitude. A gap in logical reasoning, expertise, or experience in a particular field is certainly the most visible cause. A hydroelectric expert may find it so difficult to communicate their ideas to the layperson that they are forced — sometimes painfully — to avoid certain topics altogether. They become aware of the gap between themselves and their interlocutor through subtle clues in the conversation.

It is precisely this strategy that I intend to adopt with artificial intelligence. I want to begin by assuming that the machine "feels," that it "thinks," and then determine — through certain signs — whether there is an aptitude it lacks. This aptitude is not that of pure sensation, nor that of pure reason, where the machine imitates the human mind quite convincingly, but lies exactly between the two faculties: it is understanding, conceived as the ability to form concepts from the sensible world.

If we adhere to Kantian theory, the absence of synthetic *a priori* judgements may be attributed to several factors: (a) a deficiency in pure intuitions, which we have already discussed, (b) deficiencies in pure forms, or (c) the absence of the I that enables subjective deduction — that is, the unity of all impressions, sensations, and perceptions passing through consciousness. In this chapter, I would like to examine point (b). I will reserve point (c) for the following chapter.

3.4.1 Deficiencies at the Level of the Categories

I have already discussed the possibility that the machine lacks pure intuition of space and time, which are the pure forms of sensibility, and I have shown that this hypothesis is plausible. I would now like to consider an alternative: that the machine may not be endowed with the twelve categories that serve as the pure forms of understanding. To recall, these are grouped into four classes:

1. **Quantity:** Unity, plurality, totality
2. **Modality:** Possibility, existence, necessity
3. **Quality:** Reality, negation, limitation
4. **Relation:** Substance, causality and dependence, community

In chapter 2.4, I suggested that these four types of categories roughly coincide with the four components that François Chollet identifies in the abstract reasoning involved in solving ARC tasks (the machine's IQ tests, see chapter):

1. **Arithmetic:** The ability to sort objects based on shape, colour, appearance, or movement, relying on addition, subtraction, and comparison.

2. **Object persistence:** The ability to identify and manipulate objects in such a way that a change in orientation or context does not erase the recognition of that object.

3. **Geometry and topology:** The ability to perform spatial manipulations on objects, such as rotations, orthogonal symmetries, translations, deformations, repetitions, associations, etc. In particular, it becomes possible to determine the distance between two objects.

4. **Goal-directedness:** The ability to attribute intentions or goals to certain objects.

When François Chollet identifies these criteria, he is indirectly concerned with the problem of what makes judgement possible within algorithms. The kind of judgement he has in mind involves a certain capacity for generalization and abstract thought. It is founded on a number of conceptual building blocks. These are notable in that they are not merely logical laws: they are the very glue that binds, within a single thought, a logical form to a sensible content. In simpler terms, the categories of the understanding are abstract forms that structure sensible information, and they exist only through it.

The pressing question, then, is whether current algorithms possess such structural concepts. Rather than examining the four classes of categories one by one, I would like to focus my efforts on the categories of Quantity. These are, due to their simplicity, the most likely to be found in a machine. They consist in drawing a species-level distinction between general judgements and particular judgements, such as the following:

- P1: "This is a thinking machine"
- P2: "There exists (at least) one thinking machine"
- P3: "Every machine is a thinking machine"

From a mathematical perspective, the difference between sentences P2 and P3 can be captured by introducing the symbols "there exists" and "for

all," respectively written as $\exists$ and $\forall$. These symbols allow us to express the idea that a certain property — such as being capable of thought — holds for a given machine M (a particular statement), or for the entire set of conceivable machines M (a general statement). Equipped with this vocabulary, sentences P1, P2, and P3 can be rewritten as:

(i)

(i) M is a thinking machine (a particular machine thinks)

(ii) $\forall M : M$ is a thinking machine (every machine thinks)

(iii) $\exists M : M$ is a thinking machine (there exists a thinking machine)

This expression can be made more precise and rigorous by using the notation $p(M)$ to state: "Machine M thinks"[17]. This yields:

(i) $\forall M\ p(M)$ (every machine thinks)

(ii) $\exists M\ p(M)$ (there exists a thinking machine)

These symbols must be used within a sufficiently precise language in order to be understood by a machine. One example of such a formalized language is predicate logic. Within this framework, a number of remarkable properties can be observed. For instance, "if every machine thinks," then "it is false that there exists a machine that does not think." Conversely, "if there exists a machine that thinks," then "it is false that no machine thinks." These two laws may be written as follows:

(i) $p(M) \rightarrow \neg\exists\neg p(M)$

(ii) $\forall M p(M) \rightarrow \neg\exists\neg p(M)$

(iii) $\exists p(M) \rightarrow \neg\forall\neg p(M)$

I should clarify, for readers unfamiliar with logic, that the operator $\rightarrow$ is what we call logical implication: if the sentence on the left side of the arrow is true, then the one on the right must also be true. Furthermore, the

[17] Strictly speaking, the symbol p is what we call a predicate. It is a function that takes a variable (in this case, a machine M) and assigns to it either the value *true* or the value *false*.

symbol ¬ expresses negation. When placed in front of a statement, it means that the statement is false. When it appears in front of $p(M)$, it means that machine M does not think. To be perfectly clear, $p(M)$ should thus be read as: "machine M thinks," while $\neg p(M)$ should be read as: "machine M does not think."

This is but a small example of the kinds of logical laws that can be constructed using the quantifiers $\forall$ and $\exists$. By identifying, one by one, the other laws of thought that predicate logic makes possible, and by formalizing them, we obtain a logical system comprehensible to a machine. This system allows one to distinguish general statements from particular ones.

3.4.2 General Logic vs Transcendental Logic

The remaining question—and the most philosophically significant one—is whether the machine constructed through all these logical laws is, or is not, endowed with *a priori* categories. Kant, indeed, drew a distinction between general logic, which objectively considers the laws of reasoning, and transcendental logic, which focuses on how these laws apply to the sensible contents of knowledge within a thinking subject. The question raised by artificial intelligence—framed in a jargon that might serve you well in sophisticated social gatherings—is thus the following: what are the necessary ingredients of transcendental logic?

As a first point of reflection, one may note that the human being, unlike the machine, employs general and particular statements without knowing their axiomatic rules. One can even perform quite ably in their absence, with no awareness of the symbols $\forall$ and $\exists$. This is how certain intuitive minds, with no mathematical training, manage to excel at logic. If the machine must be taught every permissible step, we, on our part, engage in logic as Monsieur Jourdain wrote prose: without realizing it!

This is why, from a Kantian perspective, these categories are *a priori* structures of thought in the human being. They exist whenever we think—even if we are unaware of them. They permeate, as it were, every content of our thought. I would like to illustrate this idea—that human thought is shaped by modality—with a striking example. If I say to you: "In the morning, the sun rises," or "In the morning, the American army lands on the beaches of Normandy," you will spontaneously associate each of these phrases with one of the two quantifiers $\forall$ or $\exists$. Unless you are old enough to carry the trauma of the Second World War, you will likely interpret: "The army landed that morning, on June 6, 1944," and not "Every morning, the army lands." In contrast, you will probably understand: "The sun rises every morning," and not "The sun rose on June 6, 1944!"

Of course, the human mind is subject to error. This is why we do not always assign the same modality as our speaker to the sentence we hear. It may also happen that we draw faulty inferences—that we betray, so to speak,

the laws of logic. This is what occurs when, led by immediate appearances, we get lost in the following line of reasoning:

1. The sun rose on June 6, 1944.
2. The American army landed on June 6, 1944.
3. Therefore, the army will land whenever the sun rises.

This correlation between two contingent facts—the rising of the sun and the arrival of the army—is (thankfully) not a logical consequence of the statements that "the army may land on a morning" and "the sun must rise every morning." And yet, it frequently happens, in more complex cases, that the human being makes this mistake. A striking example is the study that shows that countries where more chocolate is consumed are more likely to produce future Nobel laureates. You cannot infer from this that chocolate makes one smarter—otherwise, Belgium would dominate the field!

Despite our frequent errors, there remains a certain consensus about what counts as a valid or invalid argument. When you bring several human beings to the table, and give them enough time—setting aside ideological or personal conflicts (which are often the heart of the disagreement)—they generally end up converging toward a common understanding[18]. It may take several generations—scientific theories are the ruins of these interpretative wars—but one must admit a certain robustness, or stability, in the conclusions reached.

These two traits I've just mentioned—namely, that (1) error is human, but (2) truth ultimately prevails—do not appear to apply to machines (at least not the less autonomous ones ; I will discuss latter the case of advanced neural networks like ChatGPT). An algorithm programmed to reason within the modalities of "possibility" and "necessity" will not err with respect to the laws of logic, but it will never reach agreement with an algorithm programmed according to a paraconsistent logic, one that allows a thing to be both true and false at once. What would a debate look like between beings made of cables and gears, each stubbornly clinging to their logical principles? It seems that human flexibility is inseparable from the ability to err. Our thought is indeed furnished with logical structures, but they are not strict: they are merely ways of making sense of the world.

In conclusion, to endow the computer with the *a priori* categories of the understanding is not, simply and straightforwardly, to furnish it with

[18]This position was more fully articulated by Habermas, who explored the conditions of communicative rationality, or the criteria by which one might say that a debate leads to consensus. This requires, notably, two ingredients: a principle of debate that ensures a form of representativity, and a universal principle that ensures a form of objectivity and openness to the interests of all.

the rules for manipulating symbols. Such an approach would amount only to defining a general logic, not what Kant called a transcendental logic. That is, it would address the validity of reasoning *in itself*, but not the way in which data are themselves shaped by these laws. And yet it is only this second aspect—the relation of logic to the real world—that is truly at stake in the *a priori* categories. It would also seem that it is this capacity for intuitive generalization, which connects fundamental logical concepts to raw sets of information, that IQ tests or François Chollet's ARC exercises seek to measure, with varying degrees of success.

I will try to show, further along in this thesis, that the current paradigm of computer science misses this target: at times, it feeds the machine formal laws to be analytically digested like a recipe; at other times, it builds convincing correlations, but only *a posteriori*. These are the two paths of (a) expert systems, which think by way of logical laws, and (b) neural networks, which form statistical associations. I will try to demonstrate that the middle path—the path of *a priori* synthesis—is missed in both of these approaches.

3.5 Is the Machine an I?

Who has never experienced, when using ChatGPT, the strange feeling that the answers displayed on the terminal, impressive as they may be, lack personality? It is as if the machine synthesizes experiences, opinions, and reflections scattered across the infinite web of the Internet, but never truly offers anything original. Its statements are, of course, well-structured and well-informed; they are nuanced, sometimes humorous, but they remain on the surface of thought. One may feel that the algorithm provides a blurred image of reality, a close-up, a statistical sketch, missing contours, life, and character. Like a quantum wave, its productions swirl through hundreds of possibilities, oscillating between different points of view, yet never taking hold of a specific situation in the world. To paraphrase the title of Robert Musil's book, ChatGPT is *a man without qualities*; a man of pure *possibilities*.

This is why, after analysing the categories of understanding, I must now ask whether the machines we know today can have a "I". For Kant, the "I" is defined as the very fact that behind each isolated act of consciousness, it is indeed the same consciousness at work. Without an "I", in fact, and perhaps this is what happens to certain individuals suffering from schizophrenia, the contents of thought, sensation, and emotion would be experienced as isolated, disconnected experiences, not truly belonging to us. It would then be impossible, when faced with a landscape, for our eyes to wander from one element to another, to identify the shape of a mountain rising out of the mist or a ship lost on the horizon. This process would be impossible simply because it is no longer the same consciousness that observes the mountain

and the one that observes the boat. There is no possible connexion between these two contents.

More fundamentally, the absence of an "I" prevents us from seeing a boat or a mountain, because these objects themselves are composed of a diversity of fleeting impressions: points of colour, shapes, which it is impossible to connect. A consciousness without the Self would be a perfect illustration of continuous creation. At each moment, I would be, as it were, another "I"; so there would be no memory, nor any concepts. The world, if it still existed within us—one could doubt whether such a consciousness would be aware of anything—would, at best, resemble a vast kaleidoscope.

In the final part of this thesis, I will attempt to provide a phenomenological definition of the "I", influenced on this Kantian reading; and I will try to show that the "I" relies on a principle of differentiation that requires an intrinsic form of ambiguity. Indeed, in the conception of the world, there is a capacity of the mind to compare the grand fresco of sensations that befall the consciousness in such a way as to identify, within it, vast sets of objects. The boat, the mountain appear to us precisely because they are differentiated from the background on which they first appear. This implies, from a fundamental point of view, that perception is not given as a collection of isolated points that our mind connects to form categories. On the contrary, perception first presents itself as a Whole, after which we only divide it into various and distinct objects. This is why we experience a kind of vertigo, a fascination, when we are first plunged into the observation of a landscape. It seems to exceed the content of our thought. It touches us, it transports us. And it is only after this brief moment of bewilderment, which is truly the sense of the beautiful, that we immediately, unconsciously, begin to decompose this indescribable unity into as many known objects.

In contrast, the machine would find it difficult to experience this vertigo. The reason is that it is presented with sensory information in a pre-digested form. The structure of computer language requires us to tell the algorithm: "Here is an image: it is a matrix of points. Each pixel is represented by three integers defining the colour, one integer defining the brightness, another defining the albedo, and so on." In doing so, the image to be analysed has already been split into a set of minimal units. It is no longer a Whole presented to the consciousness, but isolated contents, each separate from the other.

3.6 Does the Machine Think Differently?

Before concluding this critical section on the limitations of machines, I must briefly address the last alternative I raised concerning machine judgement. This, by far the most favourable to the artificial intelligence project, consists in admitting that algorithms have the capacity to judge, both in terms of

analytical, synthetic *a posteriori* judgements and synthetic *a priori* judgements. Under this hypothesis, however, we cannot ignore the observation of differences in reasoning modes between humans and machines.

On the one hand, the machine knows how to do many things that humans cannot: it possesses, among other irrefutable qualities, a computational power and precision that are hard to rival. But on the other hand, the machine fails to identify certain categories or perform certain abstractions that are natural to our human counterparts. This occurs, for instance, when it underperforms on François Chollet's ARC tests or, more famously, when it identifies Black people as "gorillas"[19]. »

If, then, we accept that the machine is capable of the same types of judgement as humans, we must nonetheless acknowledge that it does not necessarily exercise them in the same way. Differences exist in their application, leading to just as many observable behavioral disparities. What is the exact origin of these differences? How can we pinpoint them?

A first order of difference has already been raised: it pertains, from the standpoint of understanding, to the laws of logic. I have shown that different formalisms can be used to program a machine, leading to laws that are quite foreign to each other. Some of these laws adhere to the *a priori* forms of the human mind recognized by Kant: this is the case with classical logic. But other laws may violate them, such as when one says that something can both be and not be. Such a proposition generally offends human intuition, although in many cases it can still be made sense of.

To say that the machine is capable of *a priori* synthetic judgements means nothing more than this: sensory data are shaped *a priori*. There exists a structure, an order, a prism, that intervenes between sensation and intellect. If, for us humans, points in space appear as squares, circles, or triangles, what would it be like in the eyes of a machine? Could it not possess forms of its own?

The ARC test, under this hypothesis, would not so much be a test of intelligence as a test of discrimination, akin to the one used in Philip K. Dick's short story: "Do Androids Dream of Electric Sheep? [20]" It would serve to identify, by presenting significant problems, whether one is facing a human or a cleverly disguised machine.

There is something arbitrary about the Raven matrices used in the ARC tests. While the expected response seems intuitive to us, the connexion

[19]This incident, which occurred on July 1, 2015, made headlines when Jacky Alciné and his Black friend discovered that the algorithm had categorized them as "gorillas." One might have thought the problem had been fixed. Yet, in 2023, the application still fails to identify monkeys. We will see that with neural network programming, it is indeed difficult to provide guarantees. Even when the system performs well on its test database, it is not excluded that an error could be made in a real-world case.

[20]In this story, which gave rise to Blade Runner, the robots are indistinguishable from humans, such that certain people are tasked with identifying them using an experimental device.

between the inputs and outputs is not a matter of pure logical deduction, but of simple induction. Now, induction partly depends on our thinking habits. We are accustomed to representing geometric shapes such as squares or triangles and transforming them through symmetry. We are familiar with these shapes because they are our instinctive way of structuring information. On the other hand, we are less adept at performing complex calculations. That's why we easily succeed in the ARC tests, but will likely struggle to complete the following number sequence:

$$3958585969, 11875757907, 35627273721, 106881821163...$$

Think for a moment about this problem before seeing the solution. It is unlikely, I believe, that you will find 320645463489. Yet it is merely the result of multiplying the initial number by three at each new step. This operation is complex for a human, but not for a machine. It is reasonable to think that a computer could obtain better results on this exercise than a human. It would be theoretically possible to conduct a battery of tests, such as Raven's matrices, in which only computers perform well. This would be a test aimed at determining whether the subject has the same intuition as the machine.

The points I raise here do not invalidate the central idea of my thesis: namely, that the machine does not possess the same way of structuring information. They do, however, represent an objection to the philosophically naive simplification of viewing the ARC tests as a general intelligence test. This simplification already exists extensively in the controversial world of psychometrics. Authors are known for making optimistic claims about the relevance of the IQ test, and Raven's matrices in particular, to measure individuals' intelligence. However, it is notable that the average IQ varies widely across countries and cultures. Some regions of sub-Saharan Africa report average IQ scores around 60, which corresponds to the threshold of mental retardation. Should we then conclude that people from these ethnic backgrounds are "backward" human beings—carrying all the pejorative connotations of this term? Or should we instead acknowledge that IQ depends on culture and the way individuals approach certain problems[21]? Over the last few years, *culture fair* tests have emerged, aimed at minimizing, in the formulation and selection of questions, the factors that could unfairly boost the performance of people from Northern countries. It is legitimate to wonder whether the ARC tests might contain biases, some favoring humans and others favoring machines. I will briefly mention a few of them here:

[21] In 1948, Raymond Cattell created an IQ test without verbal tasks, designed to overcome cultural biases. By the 1970s, however, studies were highlighting the presence of biases (see especially John F. Ireland, Marvin W. Kahn, *How Fair Is the Culture I.Q. Test?*, Sage Journal, V.25-1, April 1979.

- **Geometric shapes:** The exercises seem to be based on a geometric culture that favors humans, particularly those from Northern countries. The shapes to be manipulated resemble figures from Euclidean geometry: squares, triangles, etc. It might be interesting, purely speculatively, to ask whether machines have preferences when the shapes are presented in a different type of geometry?

- **Response time:** An absence of time limits, or a high time limit, is favourable to human beings. Moreover, it encourages a measured response, rather than an instinctive one. By reducing the allotted time, it is plausible that we would no longer be measuring reasoning as such, but rather immediate perception. To discriminate between humans and machines based on their perception of time, note that we could consider a simultaneity test. The goal of this test would be to determine whether two events are, or are not, simultaneous. Suppose we are confronted with a sound (the sound of a vase breaking, a creaking door, a gunshot, thunder...) and a video (a vase falling, a door opening, a bullet hitting a target, a lightning flash...). The sounds and images are presented separately. The test would ask the subject to determine when the sound is most likely to occur in the video. For example, the sound of breaking dishes should be heard when the vase hits the floor, while the sound of thunder should follow a few seconds after the flash of lightning. The human mind is accustomed to synchronizing visual and auditory stimuli. In contrast, this would not be an easy test for the machine, as it would need to understand what is represented in the video.

- textbfPixels: The exercises are presented in pixel form, which is advantageous for the machine. One might wonder what the performance would be if the shapes to be analysed were continuous forms? In that case, the protocol would need to be refined to determine, with as little ambiguity as possible, which responses are acceptable. This could be achieved either by constructing neural networks to define the possible responses, by consulting a human jury, or by hybridizing the two approaches.

What to take away from this chapter

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

- A *machine* encodes information in the form of indivisible units (pixels, bits, etc.); it seeks to statistically associate these units to form concepts.

- A *human being*, on the contrary, apprehends information as a whole; it uses concepts to isolate objects within this whole.

It all seems as if the human mind possesses conceptual structures that alter its overall perception, connecting the parts to the whole, whereas algorithms, lacking prior concepts, merely deal with objective, although independent pieces of information.

I illustrated these functional differences using several perceptual illusions (Kanizsa, Ponzo). In the next section, the question will be raised: what do the human being's prior conceptual structures actually look like? I will take this opportunity to criticise Kant's theory.

Revision of the Concept of Judgement

Toward a Mereological Theory

Chapter 4

The Forgetting of the Understanding: Dead of Kantian Judgement

Having reached this point in our intellectual journey, it might be wise to summarize the virtues of Kant's critique in light of the question of artificial intelligence. I believe I have suggested—avoiding unnecessary technicalities—that philosophy identifies three kinds of judgement, each grounded in a different faculty of the human mind: synthetic *a posteriori* judgements pertain to Sensibility, synthetic *a priori* judgements pertain primarily to the Understanding, and analytic judgements are based on Reason. I say that synthetic *a posteriori* judgements pertain "primarily" to the Understanding, and not "strictly," because they straddle both Sensibility and Understanding. In truth, all synthetic judgements—whether *a priori* or *a posteriori*—draw upon data from the sensible world, though they do so in different ways: *a posteriori* judgements rely on experience to reach their conclusion, while *a priori* judgements are grounded in the categories of the Understanding. These do not arise before the human being has experienced numbers, triangles, and angles, yet their truth does not depend on such experience.

To exemplify these three types of judgement more concretely, we can say that synthetic *a posteriori* judgements are empirical facts, revealed to us only through the daily experience of a world that is fatty, incongruous, rough: propositions such as "this table is white" or "bronze is an alloy of copper and tin," or again "the production of CO_2 causes global warming," are derived by observing consistencies, recurrences within the order of phenomena. There is nothing in the concept of "CO_2" that allows us to deduce, with irrefragable necessity, that it causes warming. If this truth followed solely from the concepts involved, it would have sparked far less debate. Alas, it required measurements, observations, hypotheses—so much so that

uncertainty remained possible; such is the nature of contingent facts: they are always open to doubt.

Analytic judgements, on the other hand, express nothing more than what is already contained within the terms of discourse. If I say, for instance, that "Napoleon's white horse is white," I add no new information. The same holds for syllogisms : if I know that all squares are rectangles and all rectangles are quadrilaterals, I may deduce that squares are quadrilaterals. In doing so, I learn nothing new about squares that I did not already know. A deductive reasoning, no matter how learned, merely clarifies truths already latent within the axioms of the theory. It places us in the position of Meno's slave, who discovers, by following Socrates' guided questioning, logical conclusions. "These opinions," Socrates comments, "have just awakened in him like a dream. And if one questions him often and in various ways on these same things, do you not think he will ultimately have knowledge of them as exact as anyone's[1]?"

Synthetic *a priori* judgements, finally, are reasonings that reduce neither to the laws of logic nor to the results of experience. They betray the existence of fundamental structures within the human mind. As an example, Kant refers to propositions such as "5 + 7 = 12" or "the sum of the angles of a triangle is equal to 180 degrees." These, he argues, are not analytic truths. There is nothing in the concepts of "triangles" or "angles" that entails, with absolute necessity, such a relation. Nor are these mathematical truths properties attributable to objects in the world. Rather, they stem from the conceptual tools we use to understand the world: namely, the intuition of time, the intuition of space, and the categories of the Understanding. These tools, unfortunately, are not directly visible. That would be too easy. How can one see the prism through which the world appears? If you have worn distorting glasses since birth, you cannot know that the world is distorted—or even that you are wearing glasses. What you will notice, however, are constants, optical laws, shared by all the things that fall under your gaze. Objects will exhibit curves, angles, focal points, characteristic symmetries. Only afterward might you begin to suspect that these laws do not belong to the world itself, but to the structure of your lenses. This means that a human must first be in contact with the world before they can conceive of mathematical properties concerning angles, triangles, or numbers. But once they discover these geometric and arithmetic truths, they feel as though they do not depend on worldly experience. They are the *a priori* conditions of any experience; laws inherent to the structures of human thought.

I have defended the idea that it is the absence of synthetic *a priori* judgement—and not the absence of synthetic *a posteriori* or analytic judgement—that underlies the functional differences between the human mind

[1] Plato, *Meno*, trans. Cousin, Philosophie, September 2009, p. 65.

and the machine. The somewhat technical analysis I carried out of François Chollet's ARC exercises appears compatible with the idea that artificial intelligence, as currently conceived, cannot intuitively and categorically structure sensory information. What do I mean by that? That the conceptual prism of the human mind bears no resemblance to that of the machine. When we program an algorithm, we make the mistake of thinking that the spatial and temporal structures of the world—such as squares, circles, or instants—exist in the world just as we perceive them. We then aim to create a machine capable of identifying them. But if we take Kant's great insight seriously, it may be that these forms result not only from a proper analysis of the world, but also from the way information is structured within our minds.

It is hard to deny that the structure of data in human beings and in machines shows substantial differences. In algorithms, data is provided in the form of discrete, indivisible units: the pixels of an image, the time intervals of an audio signal, etc. These units are treated as if they were the minimal foundations of perception. They are endowed with intrinsic properties (colour, sound intensity...) unaffected by neighboring units. The programmer's task is to determine how these points might be processed by an algorithm in order to identify the relationships between them: are they arranged so as to form a square? Do they reveal spatial depth? The million-dollar problem of artificial intelligence, to put it bluntly, lies in understanding how to reconstruct a global vision[2]. Everything happens as if the machine possessed only fragments of perception, which it must recompose to obtain a complete view. To borrow an image, the structure of artificial perception resembles a sieve. Any beam of light that passes through this sieve appears as particles scattered randomly across a blank page (see figure 4.1).

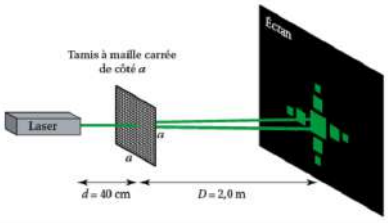


Figure 4.1: The *a priori* structure of machine perception may be likened to a sieve. All information is discretized and represented as a grid of points independent from one another.

In human beings, by contrast, perception may not consist of indivisible units. The segmentation of reality is not comparable to digital processing,

[2]The total sum of prizes awarded by François Chollet and his associates for successfully completing the ARC tests is indeed 1.1 million dollars. These tests, as we have seen, require generalizing structures from a very small number of examples.

which extracts information from an image, a sound, or a perception by discretizing it into small intervals of time, of distance, and assigning numerical values to each of these segments. For humans, information is structured in a more holistic and processual fashion. The mind attends to the contrasts within the perception of an image or a sound. It segments the whole of perception by using familiar tools (objects, causality, geometric figures...). It is because we already possess the category of the triangle that data appears to us in the form of a triangle; and only later, through meticulous analysis, does that triangle reveal itself as a mere cloud of points. We first perceive key shapes; only afterward do we notice the imperfections. Moreover, there appears to be no definite limit to the process of decomposition. We may divide a perception by isolating differences and contrasts within it, and then return to perceiving the whole. In other words, I wish to put forth the hypothesis that human perception operates through a movement back and forth, descending from the whole to the parts, and ascending from the parts back to the whole.

This dual descending and ascending process is made possible, according to Kantian theory, by the pure intuition of space, the pure intuition of time, and the categories of the Understanding. These forms are, to some extent, scientifically verifiable—or at least compatible with certain scientific findings. In neuroscience, the Marburg theory seeks to identify, within human experience, the existence of structures that determine perception *a priori*. This line of thought originates with Cassirer[3]. It gave rise, among other things, to the theory of the "Kantian Brain[4]":

> *The "Kantian Brain" model merges the philosophy of Immanuel Kant with modern brain research, proposing that the brain actively generates predictions and interpretations based on prior knowledge. It emphasizes imagination, creativity, and spontaneity in perception and cognition.*

It is, however, uncertain whether the way these forms are defined by Kant adequately reflects the modern progress of science. The pure intuition of space, in particular, falls within a Euclidean vision of the world, whereas physics now works with non-classical geometries, in which the shortest distance between two points is not a straight line. Furthermore, the categories of the Understanding convey a universalist view of the human mind, where each individual is born endowed with the same logical structures. Yet studies in anthropology seem to suggest that there exist cultures unfamiliar with the concept of number. They use only the categories "one", "two", and "many".

[3] Katherina Kinzel, *Neo-Kantian Conceptualism: Between Scientific Experience and Everyday Perception*, British Journal for the History of Philosophy, 08 Aug 2023.

[4] Valery Yevarousky, *The "Kantian Brain" Concept: Modeling the Limit of Human Perception and Existence*, September 2023.

More radically still, traces have been found in India of rather sophisticated logical systems admitting that a thing may both be and not be. One need not even travel to the banks of the Ganges to make such a discovery. In Hellenistic Greece, the Megarians were already disputing Aristotle's law of the excluded middle, claiming that a thing is either true or false, but that there exists no in-between[5].

It is therefore perhaps wise (1) to critique Kant's *a priori* forms and (2) to redefine them in such a way that their expression might be more compatible with the evolution of modern science — and of computer science in particular. To this end, we must first understand why, despite the renown Kant's theory enjoyed for over a century, synthetic *a priori* judgements — and more specifically the faculty of the Understanding — seem to have gradually vanished from the contemporary scientific conception of the world? One might also ask how this gradual historical transformation gave rise to the epistemological structures used by today's algorithms?

While there certainly still existed neo-Kantians at the beginning of the twentieth century, particularly in the context of discussions with Niels Bohr on the interpretations of quantum physics[6]; and while the notion of *a priori* faculties remains quite operational even today in neuroscience — for example, when identifying the brain areas related to the spatial organization of human representations — these echoes of the Understanding remain but anecdotal phenomena. It must be acknowledged that the existence of pure forms and universals frightens our timid relativism, more concerned with understanding how Man (and his observations) is the measure of this or that thing than with identifying universal constants.

In this section, I would like to argue that one of the most significant factors behind what I shall call "the forgetting of the faculty of the Understanding" — or, more simply, "the forgetting of the Beautiful" — is to be found in the very structure of mathematics. As it happens, Kant's two favored examples for illustrating the existence of synthetic *a priori* judgements are drawn from arithmetic and geometry. Kant writes on several occasions, in the introduction to the *Critique of Pure Reason* and in the Transcendental Aesthetic, that "5 + 7 = 12" and that "the sum of the angles of a triangle is 180 degrees." As I have shown, these examples are already present in Plato's *Theaetetus*. They have served as scaffolding, or as reference points, for the evolution of the philosophical concept of judgement throughout history. Yet, around the end of the 19$^{\text{th}}$ century and the beginning of the 20$^{\text{th}}$, groundbreaking results began to emerge that challenged the truth of these very propositions.

[5] On this topic, see the excellent work by Łukasiewicz, *Du principe de contradiction chez Aristote*, 1910, French trans. by Dorota Sikora, preface by Roger Pouivet, Paris, L'Éclat, 2000

[6] See especially the chapter entitled *Bohr's Complementarity and Kant's Epistemology*, from the volume *Niels Bohr, 1913–2013*.

On the one hand, the revision of Euclid's fifth postulate led to the emergence of non-Euclidean geometries. These non-intuitive theories — within which the sum of the angles of a triangle is sometimes more than 180 degrees, sometimes less — raise questions about the origin of geometric truths: are they conventions? results born of experience in the world? mere contingencies of our human mind? The mathematician Poincaré wondered what our geometry would have looked like had we been born not on solid ground, but in an aquatic environment, where bodies deform and light refracts.

On the other hand, the mathematician Hilbert proposed a project to axiomatize mathematics — in particular arithmetic — so that additions such as "5 + 7 = 12" could be demonstrated from a small number of axioms. In both cases, the idea gained ground that mathematical truths are logical truths, analytic, deducible by reason alone. The notion of intuition, though still present, takes on a different meaning. It is no longer the *a priori* structure, the scaffolding without which conscious phenomena could not exist; it becomes a certain aptitude to grasp the self-evidence of an axiom. Far, then, from the idea that the world is given through the intuitions of the human mind — by conforming to their form — one admits instead that intuition is a tool to grasp the first truths of a world that exists outside us — and of which we may apprehend the true nature.

In addition to turning Kantian intuition on its head, these developments in the field of mathematics have led to a new — not necessarily fatal, but somewhat different — understanding of the notion of judgement. With the rise of formalism and logicism, judgement becomes once again, as in Plato's time, an instrument of reason. Yet far from seeking to grasp the nature of universal essences, its role is confined to calculating, deducing, formulating. This calculating reason, made entirely of numbers, manipulates symbols, aligns them, transforms them, erases them when necessary, producing new sequences. Human thought is no longer the expression of *a priori* forms; it is a clever machinery of infinite tapes, reading heads, and logical laws signed "Alan Turing." For the first time, it takes the form of a calculating machine. A machine, to be sure, of a Dantean nature — dazzling in its speed and precision — that seems capable of answering questions and thinking autonomously.

In the sections that follow, I will focus more specifically on the way non-Euclidean geometries have challenged the idea of a pure intuition of space. The goal will be to determine whether Kant's canonical example — that "the sum of the angles of a triangle is 180 degrees" — can still be considered a synthetic *a priori* judgement. In the following part of this work, I will focus more specifically on the arithmetic proposition that "the sum of 7 and 5 is 12." I will take the opportunity to show how this paves the way for the emergence of that now-indispensable tool: the computer.

4.1 The Crisis of the Foundations of Mathematics

Between the end of the 19$^{\text{th}}$ and the beginning of the 20$^{\text{th}}$ century, a more or less fundamental reevaluation of the ontological and epistemological status of mathematics took place. Though it was not always perceived as such at the time, this reevaluation had such a profound impact on the later development of science—especially on disciplines such as logic, set theory, and category theory—that it is now universally referred to as a "crisis."

Some trace the origins of this phenomenon to around 1870,[7] when the philosophical debate surrounding non-Euclidean geometries reached its climax. Others prefer the year 1902, in reference to the letter Russell sent to Frege, informing him of the existence of a logical paradox undermining the foundations of arithmetic.[8] But if the inauguration of this event is itself a matter of debate, what can be said of its conclusion? Some believe the crisis ended shortly after the publication of Gödel's theorems,[9] while others consider its consequences to have extended far beyond. What remains certain is that this episode spilled much ink and deserves our full attention.

Several controversies lay at the heart of the crisis. The most widely known today is undoubtedly that surrounding the many paradoxes of set theory—upon which one strove, with varying success, to anchor arithmetic. These include the paradoxes of Russell, Richard, Burali-Forti, and Cantor.[10] A librarian by the name of Berry distilled the paradoxes' very essence into a single contradictory object, defined as: "The smallest natural number not describable in fifteen words or fewer." The trick, of course, lies in the fact that the quoted sentence consists of precisely fifteen words.

These logical paradoxes left a lasting imprint on posterity but went almost unnoticed by the mathematicians of the time. Regarded as mere "puzzles" or "language games," they managed only to provoke the irony of the renowned mathematician Henri Poincaré: "Logic is no longer sterile—it gives birth to contradictions."[11] These paradoxes are said to have been resolved with the near-simultaneous publication of *Set Theory as Based on the Theory of Types* (Russell) and *Investigations in the Foundations of Set Theory I* (Zermelo, 1908).[12] It was only years later that the full depth of their implications became clear, particularly regarding the project of building mathematics solely upon analytic judgements.

The second major controversy that left its mark on history concerns the status of mathematics and harks back to the quarrel between formal-

[7] Lepage and Fradet [114], *La crise des fondements : quelle crise ?*, see Avant-propos, pp. 1–5.
[8] Barot, *En quoi la crise des fondements des mathématiques est-elle terminée ?*, p. 1.
[9] *Ibidem.*
[10] Lepage [114], *op. cit.*, p. 1.
[11] Poincaré, *Les mathématiques et la logique*, p. 306.
[12] Lepage, *loc. cit.*

ism and intuitionism, embodied respectively by Hilbert and Brouwer. The former holds that mathematical truths correspond to objects of our intuition—namely, symbols—but that this intuition must unfold through an axiomatized system of rules. From this perspective, formal mathematics must be accompanied by a metamathematics—preferably finitist—whose role is to constructively prove the validity of the inferential rules employed by mathematics proper.

Brouwer, by contrast, defends a vision in which no formal language, however rigorous, can validly express mathematical intuition—which is *a priori*. He rejects the idea that the existence of a mathematical object is equivalent to its non-contradiction, accepting only constructive proofs. For the same reason, he excludes the principle of the excluded middle from the rules of inference. Upon closer inspection, both these positions appear to emerge from the ruins of the Kantian system (we shall later see what undermined it[13]). Both offer innovative responses to the new status of mathematics.

A third controversy, more discreet this time, centres on the Polish Lvov–Warsaw school and the thought of Stanisław Leśniewski (1886–1939). This philosopher and logician, working in parallel with Zermelo–Fraenkel, opened a third path for founding mathematics. His system is composed of an *Ontology*, a *Protothetic*, and a *Mereology*,[14] which respectively offer alternatives to propositional calculus, predicate calculus, and set theory. Leśniewski belongs to the logicist tradition, but whereas Frege limited himself to restricting the use of his Basic Law V, and Russell introduced the theory of types, Leśniewski sought to define the notion of a "class" in a way that was non-contradictory and more faithful to intuition. His major innovation lies in envisioning open systems, in which the terms of the language (names) are not defined once and for all but can be indefinitely added. The debate revolved around determining the precise relationship between the Pole's solution and those of his contemporaries.

Rather than merely surveying these three controversies one after another, or delving deeply into a single one, I would like to propose a new interpretive perspective: that throughout the tumultuous period known as the foundations crisis, a deeper conflict was unfolding—one between two visions of mathematical truth that reach back to the very origins of philosophy.

On the one hand, there are the unwavering defenders of the synthetic *a priori* judgement who, despite the difficulties encountered by Kantian theory, strive to preserve aprioricity either (a) in the form of pure intuition or (b) through a mathematical Platonism. These include the intuitionists and the formalists.

On the other hand, there are those who, renouncing *a priori* intuition,

[13] See section 4.4, *Objections to Kant*.

[14] Mereology was Leśniewski's first system. It is axiomatized but not formalized. The Ontology and Protothetic were later added to provide more solid foundations.

turn either (c) to the side of *a posteriori* judgement, or (d) to that of analytic judgements. Among these we find the constructivists, the empiricists, and finally, the logicists.

To properly bring these underlying tensions to light, it may be useful to begin by synthesizing what Kant and Plato contributed to the epistemology of mathematics. What is a mathematical object? What is a correct piece of reasoning? Upon close consideration, Kant and Plato have this in common: both are, to a large extent, motivated by a fascination with the power of mathematics—and a desire to account for it.[15] What do we mean by this?

There is in the practice of mathematics something akin to the revelation of a truth which, though not given to reason in advance, imposes itself with utmost necessity. What once seemed self-evident may be overturned by the laws of logic alone; what once appeared doubtful may suddenly receive the sanction of demonstration. On one side, the mind feels free to craft its own concepts, independent of the constraints of the sensible world, and to imbue them with whatever it pleases. On the other, it continually senses that these very concepts surpass it, that they cannot be entirely penetrated, that they obey laws of their own.

The mathematical object often appears to possess the same autonomy as a physical object—except that its contours are drawn in numbers, its form is one of essences, and its matter, universal. Whoever engages with it is thus compelled to answer the thorny question: *how can we account for the apparent autonomy of mathematical objects while preserving, at the same time, the apparent autonomy of reason itself?*

Euclid's *Elements* served as the model for mathematics for several centuries.[16] This monumental edifice revealed to the world that it is possible to discover the properties of geometric figures by reasoning directly from a set of definitions and axioms. Thus, mathematics could be freed from its sensible representations and ushered into a logic of essences. In doing so, its proofs attained universal generality: they became true of every circle, every triangle—not merely those that had been drawn. This leap was invaluable. It opened the vision of an entire realm of truth, still awaiting conquest by

[15] Was not the portico of the Academy inscribed with the words: "Let no one ignorant of geometry enter here"?

[16] Though clearly the most influential of its time, this work—of uncertain date—is far from the earliest example of an apodictic presentation of mathematical results. Shortly before, a scholar like Autolycus of Pitane composed a treatise on rotating spheres, consisting of definitions and twelve propositions. These results pale in comparison to the 470 propositions in Euclid's *Elements*, but suggest that this style of mathematical exposition belongs to a more ancient tradition. Moreover, the *Elements* themselves were a compilation of earlier works: those of Hippocrates of Chios (Books I–IV and XI), Pythagorean arithmetic (Books VII–IX), the theories of Eudoxus (V–VI and XII), and Theaetetus' results on quadratic irrationals (Book X) and on polygons (Book XIII). For more, see Chapters 1–2 of Rosenfeld [166], *An History of Non-Euclidean Geometry: Evolution of the Concept of a Geometric Space*.

the intellect.

Without the publication of Euclid's *Elements* and the development of mathematics, perhaps Plato would have had little to object to in Protagoras's relativism.[17] His system seems to have been forged to justify: 1) the superiority of essentialist judgement over visible appearances; and 2) the autonomy of mathematical objects, which are none other than the Ideas. In doing so, Plato responds to the question raised by mathematical experience. He shows that the laws of reason possess a certainty absent from sensible knowledge. He further shows that the intelligible confronts reason and precedes it. He even goes so far as to reduce all of reality to these pure Forms. With him, the realist position in mathematics is born—mathematical Platonism, named in his honour.

Kant's reversal, as we have seen, consists in returning mathematical truths to the domain of sensible intuition, via synthetic *a priori* judgements: "Mathematical judgements are all synthetic."[18] In one stroke, he deprives reason of the freedom to create concepts in the void, while preserving the universal nature of mathematical propositions. As for the mathematical objects themselves, they retain, of course, their autonomy and internal consistency. But they now contain only what the understanding has put into them. Rather than being emissaries from some foreign realm, they reside within the subject's own faculties.

Thus, with Kant, the object of mathematics becomes the discovery of the *a priori* laws governing our spatio-temporal intuitions. "The brilliant example of mathematics shows us how far we can advance in *a priori* knowledge without the aid of experience. It is true that it concerns only objects and cognitions that can be represented in intuition; but this circumstance is easily overlooked, since the intuition in question can itself be given *a priori*, and thus scarcely differs from a mere pure concept."[19]

Whereas Plato's mathematical object is given timelessly, the same does not hold for Kant. For him, mathematical existence—at least in its phenomenal aspect (since knowledge cannot access the thing-in-itself)—depends on the possibility of mental construction. Only what can be constructed in intuition and in time, through a succession of finite operations, is said to exist.[20] Non-constructive existence proofs, such as the assertion that every

[17] As indicated in section **??**, the Theory of Ideas owes much to the Pythagoreans. It is known that they placed at the heart of their teaching a philosophy and mysticism of number, which in turn refers to a pure science of geometry. This supports the idea that mathematics contributed significantly to Plato's epistemology.

[18] Kant [99], *Critique of Pure Reason*, Introduction V, p. 58.

[19] Kant [99], *op. cit.*, p. 52.

[20] In Kant's theory of schematism, each step involved in the construction must be finite, but the process itself may be indefinite. This notion of potential infinity, I believe, is central to Kant. It helps explain the antinomies of reason, particularly regarding whether the world is finite or infinite. There is, in effect, this dual idea that the limits of the world endlessly recede before reason, and that they must nonetheless be found.

ordinary differential equation admits a solution, would likely not qualify as mathematical objects in Kant's eyes. This stance laid the groundwork for a major epistemological doctrine: intuitionism.

What unites both perspectives is a shared faith in the certainty of the *ad geometrico* method of demonstration. For Kant as for Plato, the mind, when it reasons rightly, can attain truth with indubitable clarity. For the one, correct reasoning lies in detachment from the sensible; for the other, in its proper ordering. Yet it is important to note that this certainty, already by the 18$^{\text{th}}$ century, was no longer absolute. With the exception of Euclidean geometry, mathematical disciplines suffered from a lack of rigor and foundational clarity.[21] There existed no axiomatics for arithmetic or analysis. Moreover, rules of reasoning were not always clearly defined. Finite logical deductions were blended with infinitesimal elements. Potential infinity lurked behind every passage to the limit. Yet despite these imperfections, all mathematical results were granted the same prestige as Euclid's *Elements*. At least, they were—until geometry itself began to tremble. It is this intellectual earthquake that we shall explore in the pages that follow.

4.2 The Book of God and the Parallel Worlds

There are certain books whose purity, elegance, and depth have left such a profound mark on the history of humanity that, were their author unknown, one might easily take them for a book of God. Among these exceptional works stand, without the slightest doubt, Euclid's *Elements*. The number of editions of this mathematical treatise—originally composed of thirteen distinct books—is surpassed only by the Bible.

But what is shaken by the mathematical advances of the 19$^{\text{th}}$ century is precisely the all-powerful authority of this mathematical Bible. Let us note well that this disillusionment did not occur overnight. Since the time of the Greeks, one of Euclid's postulates—known as the parallel postulate—had already drawn attention for its lack of self-evidence. It seemed less certain, less intuitive. Many ancient mathematicians, such as Proclus, attempted to derive it from the first four axioms, but to no avail. Schopenhauer proposed to regard it as a synthetic *a priori* judgement. Yet it soon became clear that this postulate was neither demonstrable nor even necessary, and that other geometries, each more astonishing than the last, could be born from its abandonment. It became apparent that reasoning could lead to geometric truths that transcend intuition.

The parallel postulate appears in Book I of Euclid. It is stated after a sequence of 35 definitions concerning points, lines, and surfaces. It accompanies four other postulates, of a far more "intuitive" nature. I provide

[21] Kline [108], *Foundations of Mathematics*, pp. 200–201.

them here as Euclid originally formulated them[22], though they have since been expressed in much more formal terms[23]:

1. One can draw a unique straight line between any two points.
2. A segment can be continuously extended into a straight line.
3. A circle can be drawn with any centre and any radius.
4. All right angles are equal to one another.
5. Given any point A in the plane and any line D, there exists one and only one line parallel to D passing through A[24]

According to the Russian historian Boris Rosenfeld, Euclid may have introduced the fifth postulate in order to prevent a series of errors prevalent in his time[25]. It would thus have been a genuine innovation on his part. More importantly still, this would suggest that Euclid may have realized that this seemingly simple postulate was not a logical consequence of the preceding four. Indeed, everything suggests that Euclid tried to prove as many theorems as possible without making use of it. Only from Proposition 29 onward does he find himself compelled to invoke it. That proposition states (see figure **??**):

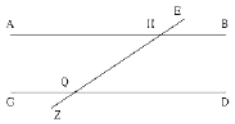


Figure 4.2: Proposition XXIX

Proposition 1 (Proposition XXIX (Book II))**.** *A straight line falling on two parallel lines makes the alternate angles equal to one another, the exterior*

[22] Following Peyrad's translation from Perrin [**?**], *Les géométries non euclidiennes*, p.1.

[23] In fact, these five postulates are not sufficient for Euclidean geometry. Today, their number has been extended to twelve, cf. Skyler, pp. 8–10.

[24] This intuitive formulation of the fifth postulate is not the one found in Euclid's original text, but rather comes from Proclus. In fact, Proclus does not even state it explicitly, but assumes it within a demonstration. Euclid's original wording is as follows: "If a line falling on two lines makes the interior angles on the same side less than two right angles, the two lines, if extended indefinitely, meet on that side on which the angles are less than two right angles."

[25] Rosenfeld [166], *op. cit.*, p. 38. His argument draws on a reading of Aristotle: "In discussing the error known as *petitio principii*, Aristotle observes: 'This is what happens with those who think they are describing parallel lines, while they are unconsciously assuming things that cannot be demonstrated in the absence of the concept of parallel lines'" (quoted from the *Prior Analytics* by Rosenfeld, *op. cit.*, p. 38).

angle equal to the interior and opposite angle on the same side, and the interior angles on the same side together equal to two right angles

From that point on, the postulate appears with increasing frequency, notably in the demonstration of the following properties:

- The relation of parallelism is transitive: if a line D_1 is parallel to a line D_2, and D_2 is parallel to a line D_3, then D_1 is parallel to D_3;
- If lines D and D' are perpendicular, then D is also perpendicular to all lines parallel to D';
- The sum of the angles in a triangle is equal to 180°;

The idea of grounding such properties on a postulate that itself seems more like a theorem than an "unquestionable self-evidence" greatly troubled geometers. Until the 19$^{\text{th}}$ century, numerous mathematicians from antiquity[26], from the Arab world[27], and from Renaissance Europe[28] tried in vain to prove the postulate[29]. Most of these efforts, when they do not mask a *petitio principii*, fall into two categories: either a new definition of parallelism is introduced, or a postulate equivalent to Euclid's is proposed in its place—though deemed more "intuitive." One example is Posidonius's definition, according to which a line D' parallel to a line D is *the locus of points equidistant from D and lying on the same side of it.*

The matter might have ended there had it not been for a man named Saccheri. This Italian mathematician, described as a precocious child and ordained a priest in 1694, became obsessed with providing a definitive proof of the postulate. Instead of seeking a direct demonstration, he cleverly circumvented the problem. He assumed the postulate to be false and tried to derive a contradiction with the other postulates (the second, in particular). This strategy is known as *reductio ad absurdum.* It led him to construct "squares" with curved sides, or triangles whose angle sums differ from 180°[30]. Saccheri failed to obtain the contradictions he sought. It was as though logic itself refused to cooperate. He died two months after publishing his work, in near-total obscurity.

[26] Archimedes, Posidonius, Ptolemy, Proclus...

[27] Al-Jawhari, Thabit Ibn Qurra, Ibn al-Haytham, Khayyam, Husam al-Din al-Salar, Nasir al-Din al-Tusi, al-Hanafi, al-Abhari, al-Maghribi...

[28] Gerson and Alfonso, Clavius, Cataldi, Giordano, Wallis, Saccheri, Clairaut, Lambert, Bertrand, Legendre, Farkas Bolyai...

[29] For a complete history of these attempts, see Rosenfeld, pp. 41–109.

[30] Rosenfeld [166], *op. cit.*, pp. 98–99.

Saccheri was not the first to reason by reductio ad absurdum. Similar ideas are found in the Arabic works of Omar Khayyam[31]. In truth, suspending Euclid's postulate, though counter-intuitive, is far from absurd. It can lead to two distinct types of logically coherent geometries: the first is known as elliptic geometry, in which the sum of the angles in a triangle is strictly greater than 180°, and in which no parallel lines exist; the second is hyperbolic geometry, where the sum of the angles in a triangle is strictly less than 180°, and where, for any line D and any point not on D, an infinite number of lines parallel to D can pass through the point (see figure **??**). It was not until the mid-19$^{\text{th}}$ century that mathematicians such as Beltrami, Lambert, Lobachevsky, Bolyai, Riemann, Gauss, and Klein[32] embraced the possibility of these theories and formalized their concepts.

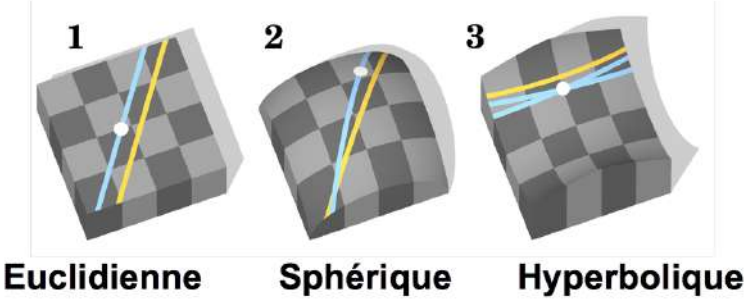


Figure 4.3: Let D (in green) be a straight line and A (in white) a point. In Euclidean geometry, there exists a unique line parallel to D passing through A. In elliptic geometry, there is none; in hyperbolic geometry, infinitely many.

All non-Euclidean geometries depend on how the length of a curve is defined, for these definitions may imply that the shortest path is not given by the "straight line" in the Euclidean sense. But insofar as the rejection of the parallel postulate—far from leading to absurdities—allows for the definition of theories as coherent as they are original, it seems legitimate to ask how many non-Euclidean geometries are possible. To guide us in this question, we may apply a version of Lie's theorem, which states that the number of possible geometries is always finite, provided the following conditions are met[33]:

- Space is n-dimensional.

[31] Khayyam Omar, *Discussion of Difficulties in Euclid*

[32] *Ibidem*, pp. 289–297 for elliptic geometries and pp. 206–221 for hyperbolic geometry.

[33] This popularized application of Lie's theorem is presented by Poincaré [152] in *La Science et l'Hypothèse*, p. 63.

- The motion of a rigid figure is possible.
- p conditions are needed to determine the position of the figure in space.

Hypothesis 2 states that a solid moving through space must undergo no deformation during its movement, regardless of its trajectory. By relaxing this condition, it becomes easy to construct an infinity of geometries, as Riemann demonstrated[34]. Today, we distinguish between algebraic, differential, and analytic geometries. To trace the impact these geometries have had on the epistemology of mathematics and on the status of the synthetic *a priori* judgement, I shall confine myself to a brief description of the two great classes of geometries that shaped the mid-19$^{\text{th}}$ century: elliptic and hyperbolic geometries.

4.2.1 Spherical and Elliptic Geometries

Let us consider a sphere and draw a plane through its center. The intersection defines what is known as a *great circle*, which may be identified with the "straight line" of Euclidean geometry. The "segments" are the arcs of these great circles. We may use them to draw triangles, squares, and so on. The geometry thus obtained is known as spherical geometry.

It is easy to see that between two non-antipodal points A and B, only one great circle can pass. But in the particular case where A and B are antipodal, it becomes possible to draw infinitely many. This result violates the first axiom. Spherical geometry is therefore not Euclidean.

The third axiom, too, encounters limitations: while it remains possible to draw circles from any center, it is inadmissible to assign them a radius greater than that of the sphere. But the more interesting question, for our purposes, concerns the fate of the fifth postulate. In the form Euclid gave it, the postulate remains entirely valid. And with good reason: all lines intersect on a sphere. But if we consider the form proposed by Proclus, the situation is quite different. This version, let us recall, states that there is only one straight line passing through a given point that is parallel to a given line. Since there are, in fact, no parallels in spherical geometry, the axiom must be reformulated as follows:

Postulat 1 (Proposition XXIX (Book II))**.** *For any point A in the plane and any line D, there exists no line parallel to D passing through A.*

Spherical geometries (*sphericae*) were the subject of many reflections and treatises throughout Antiquity. The principal motivation behind these studies was the observation of celestial bodies and the development of astronomy. In his *Almagest*, Ptolemy recorded a great number of observations

[34] *Ibidem.*

made by the Babylonians in the 8th century BCE. It is to them that we owe the division of the ecliptic into twelve zodiac signs, each spanning 30°, as well as the sexagesimal division of degrees into minutes and seconds[35]. It was not until the 4th century BCE that the Greeks came into contact with this astronomical tradition and began to develop it further.

More general properties of spherical triangles were proved in the 16th century. In an article published in 1629, the Dutch mathematician Albert Girard (1595–1632) demonstrated that the greater the surface area of a spherical triangle, the greater the sum of its angles[36]. More precisely, he showed that the difference between the sum of the angles of the triangle and 180°— a quantity known as the *angular excess* — is proportional to the area of the triangle. He even managed to extend this result to any kind of spherical polygon. He was aware that his proofs were not entirely rigorous, as they required the use of infinitesimal calculus. Euler would later provide a complete, simple, and elegant demonstration[37]. He also gave trigonometric expressions relating the area of a triangle to the length of its sides, contributing significantly to the development of the spherical trigonometry of Ptolemy and Copernicus.

All of these geometries are perfectly compatible with our intuition. Indeed, they deal with the properties of polygons situated on the surface of a sphere, which can easily be visualized within Euclidean space. It might be interesting to ask whether three-dimensional spherical geometries exist — or whether there are ones that satisfy the first four of Euclid's postulates, but not the parallel postulate[38]. These questions are difficult to answer using our spatial intuition alone, but they may be approached through a hypothetico-deductive method.

The mathematician Riemann was the first to undertake such a method, around 1856[39]. The result was the construction of a vast set of logically justified mathematical propositions, now known as elliptic geometry. This theory admits a number of models — or, to put it differently, spatial representations. In each of these models, the theoretical objects are mapped to intuitive geometric objects (lines, curves, spheres...) and governed by rules describing their behavior.

The simplest model consists in picturing a spherical lens in which each pair of antipodes represents a "point" of our geometry. Put otherwise, only one hemisphere of the sphere is retained: the second is its mirror image. The "lines" in this model are the great circles of the sphere (i.e., the circles

[35] Almageste [**?**], p. 166
[36] *Ibidem*, p. 27–28.
[37] *Ibidem*, p. 32.
[38] Let us note that in the contemporary axiomatic formulation of Euclidean geometry, there are not five postulates but twelve. The twelfth corresponds to the parallel postulate. It is the only one not satisfied by elliptic geometry.
[39] Rosenfeld [166], *op. cit.*, p. 289

of maximal diameter). If one considers any point M on the sphere and a "line" D, one observes that there exists no parallel to D passing through M. Moreover, the result stating that the sum of the angles of a triangle is greater than 180°is preserved. This model is called *spherical elliptic geometry*, to distinguish it from two-dimensional spherical geometry[40].

4.2.2 Hyperbolic Geometries

Shortly before the development of elliptic geometries, an intriguing alternative to the parallel postulate had already been glimpsed by Lobachevsky, Gauss, and Bolyai. This alternative stems from the following observation: in elliptic geometries, the sum of the angles in a triangle exceeds 180°. In Euclidean geometry, it is exactly 180°. Might there not exist a geometry in which this sum is *less* than 180°? In a paper entitled *On the Principles of Geometry*, Lobachevsky articulates this idea:

> *"It remains to assume that this sum is either $= \pi$ or $< \pi$. Both are conceivable without contradiction, and this gives rise to two geometries: the one, hitherto preferred due to its simplicity, indeed accords with all measurements; the other, which is imaginary and, by virtue of its greater generality, renders calculations more complex, admits a dependence between the size of a triangle's sides and the sum of its angles*[41]*. "*

One of the key properties of this imaginary geometry is that—much like what is observed in elliptic space—the sum of a triangle's angles deviates increasingly from 180°as the length of its sides increases. In this case, of course, the deviation is negative. That said, many other properties—less "intuitive" still—can be discerned. Among them, let us highlight the most significant[42]:

- **Limit circles**: A circle whose radius tends toward infinity does not behave like a straight line, but rather like a curve of genus 2 (that is, a two-dimensional curve), specific to Lobachevsky's geometry, and known as a *limit circle*.

- **Limit spheres**: Similarly, a sphere whose radius tends toward infinity begins to resemble a genus-3 object known as a *limit sphere*. When

[40] Note that in spherical geometry, it is possible to draw infinitely many lines through two antipodal points on a sphere, which contradicts Euclid's first postulate. In spherical elliptic geometry, however, antipodal points are identified, which allows us to circumvent this situation and preserve the first postulate.

[41] Rosenfeld [166], *op. cit.*, p. 207

[42] The four properties are drawn from Lobachevsky's article *On the Principles of Geometry*, as cited in Rosenfeld [166], p. 207.

such a sphere is intersected by a plane, the result is either a circle or a limit circle.

- **"Lines"**: The geometry within a limit sphere is "analogous" to Euclidean geometry in the plane. All the limit circles of the limit sphere can be interpreted as the "lines" of Euclidean space[43]. Furthermore, the angles between these "lines" can be measured by considering the angles between the planes in which these limit circles lie.

- **Multiple parallels**: The fifth postulate of Euclid is violated within a limit sphere. Given any "line" D and any point P outside of D, there exists an infinity of "lines" passing through P and parallel to D.

Though these properties may at first appear difficult to visualize, it is precisely because they resist our spatial intuition. Yet, it is possible to give them visual interpretations—either by projecting this space into a circle, or by projection onto an infinite half-plane. In his popular science work *Science and Hypothesis*, Poincaré offers the following analogy:

> *"Let us suppose a world enclosed within a great sphere and governed by the following laws: temperature within this world is not uniform; it reaches its maximum at the centre and decreases progressively as one moves outward, vanishing into absolute zero upon reaching the enclosing sphere. Let us define more precisely the law governing this variation of temperature. Let R be the radius of the boundary sphere, and r the distance of a given point from its center. Absolute temperature shall then be proportional to R^2-r^2. A moving object will then become smaller and smaller as it approaches the limit circle. Let us first note that, though this world is spatially bounded in terms of our usual geometry, it will appear infinite to its inhabitants. For as they attempt to approach the limit circle, they grow colder and ever smaller. Their steps grow correspondingly shorter, so that they can never reach that limit circle*[44]*. "*

The inhabitants of such a world cannot perceive the change in their own size as they traverse space, for their measuring instruments shrink along with them. It is possible to demonstrate that, for them, the shortest path from one point to another is not a straight line, but a curve. An individual beginning at point A, situated near the surface of their spherical world, and aiming to reach another point B, also near the surface, would need to follow

[43] They satisfy all of Euclid's postulates, save for the fifth.
[44] Poincaré [152], *Science and Hypothesis*, p. 84–85.

an arc of a circle. These arcs correspond to the straight lines of classical geometry. They are called *limit circles*.

If we project this space onto a disk, we obtain what is known as the Poincaré disk model. In the figure (cf. **??**), three lines—blue, red, and green—are shown, each representing the shortest path between the two points it connects. The term *limit* circle now takes on its full meaning, as these circles intersect the boundary of the disk. And yet, according to the description of this world, it is impossible in principle to reach that boundary. Moreover, we observe that through a single point, two lines (the red and the blue) may pass without intersecting a third (the green). One may thus say that the red and blue lines are two distinct parallels to the green line—both passing through the same point. The notion of parallelism, in fact, is synonymous with "non-intersection". We thus obtain a direct contradiction of Euclid's fifth postulate, which held that through a given point, only one line could be drawn parallel to a given line. A more rigorous analysis confirms that this interpretation is formally sound.

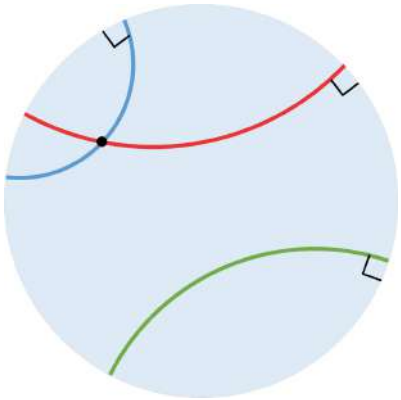


Figure 4.4: Poincaré Disk

4.3 An Imaginary Geometry

At the dawn of the 19$^{\text{th}}$ century, it was already known that the Earth was not flat — but no one would have dared to imagine that Space itself could bend beneath the stars. Picture, for a moment, the mind of an astronomer gazing through the eyepiece of his telescope. Is it not true that a straight line, drawn between two luminous points in the cosmos, pierces the void in a perfectly rectilinear fashion? The idea that such a line might veer off course, tracing the invisible folds of the sky, had not yet brushed the imagination of rationalists in dressing gowns. And yet the miracle occurred. Euclidean geometry, shedding its fifth postulate like a sandbag from a hot-air balloon, gave birth to monsters. And while our knowledge continues its

ascent toward the eternal heavens, a question drifts on every tongue: what is the relationship between non-Euclidean geometric theories and the intuitive, Euclidean space we encounter every day?

The terms chosen by Lobachevsky to describe his geometry were not selected by chance. By naming it *imaginary geometry*, he sought to suggest that the relationship it bears to Euclidean geometry is analogous to the relationship between imaginary and real numbers. Two perpendicular axes, two complementary dimensions. As long as the curvature parameter remains zero, non-Euclidean geometry behaves in a way that is almost indistinguishable from the geometry of Euclid's *Elements*. The sum of the angles of a triangle equals 180°, and through a given point, only one line can be drawn parallel to another. But once that parameter is given a positive or negative value, everything changes! Space becomes at times elliptic, at others hyperbolic; triangles possess either fewer or more than 180°, yet no longer conform to what we once believed we knew about them. The spatial structure thus conceived must be viewed as a generalization of our usual geometry. More than that: one quickly comes to realize that imaginary geometry — just like imaginary numbers — can be used to solve entire classes of real integrals that were previously insoluble.[45]

The true question that must be raised, then, is whether this geometry, with its surprising properties, might not, in a certain sense, be more real than our Euclidean space? Could it be that our space possesses this strange structure — and that we simply fail to perceive it? Lobachevsky became interested in this question very early on. He shows that for very small geometric figures, such as those with which we are most familiar, the properties of imaginary space remain imperceptible. In such a case, the sum of the angles of a triangle is virtually 180°, and light appears to follow Euclidean trajectories. The differences would only become apparent if one could conceive of very large paths or triangles.

Eager to test his hypothesis, Lobachevsky undertook a number of astronomical measurements. Referring to the data available in his time, he calculated the angles of a triangle whose first vertex coincided with the star Sirius, and whose other two vertices were antipodal points on Earth. A deviation of 0.00032 arcseconds was obtained, which led him to the skeptical conclusion that it is impossible to discover the geometry of our universe through experiment[46]. Yet the mere presence of a deviation, however minuscule, opens the possibility of a non-Euclidean space — such that only a "perfect" measurement could definitively eliminate the hypothesis.

Contemporary measuring instruments reveal that the deviation is a hundred times smaller than that observed by Lobachevsky[47]. The space he

[45] Rosenfeld [166], *op. cit.*, p. 210.
[46] *Ibidem*, p. 208.
[47] *Ibidem*.

discovered — and which today bears his name — nonetheless enjoyed a bright future in the advancement of science. Far from being dismissed as a mere curiosity or rejected on dogmatic grounds, it has been integrated as an essential element in many of the cosmologies of the 20$^{\text{th}}$ century. The Friedman–Lemaître theory, for instance, predicts successive cycles of expansion and contraction of the universe. The radius of curvature may alternately become positive — in which case the universe adopts a spherical geometry — then negative, signifying a hyperbolic one, and it may even pass subtly through zero — the special case corresponding to Euclidean geometry.

4.3.1 Lobachevsky's Constructivism

This ontological reconsideration is accompanied by a revision of the very notion of space. Until then, mathematical truths — in the Kantian tradition — were understood as synthetic *a priori* judgements. Yet Lobachevsky asserts: "Concepts must be learned through the senses — those which are innate should not be trusted[48]."

From his perspective, geometric ideas are not given a priori through the exercise of some special intuition. Quite the contrary: they are constructions of the human mind, derived from experience. In his *Geometry* (1823), he explains the process by which geometric notions are abstracted[49]. First, he says, the mind retains from its experience of solid bodies only their property of having extension. Upon this foundation it builds the idea of space, and it does so iteratively. How does it happen exactly?

The mind readily represents to itself that two extended bodies A and B can come into contact to form a third body C. This latter is conceived as a whole distinct from A and B, though inclusive of them. By repeating this experience with all bodies in nature, one may represent the entire world as a single, unified body. From this abstraction arises the idea of space. From there, the mind organises space by defining principal sections (the axes), the notion of invariant transformations (translations, rotations...), of relative positions and distance. This then allows it to define the sphere as the locus of points equidistant from a common center, the plane as the locus of points equidistant from two points, and so on.

> *Strictly speaking, all we know of nature is motion, without which no sensory impression is possible. Consequently, all other notions — such as those of Geometry — are artificially conceived by our mind, in that they are abstracted from the properties of motion: space itself, on its own, does not exist for us. That being the case, no contradiction can arise in the mind if we admit that certain forces of nature follow one particular geometry,*

[48] Quoted from *ibidem*, 209.
[49] *Ibidem*, pp. 210–212.

while others follow their own geometries[50].

4.3.2 Poincaré's Consensualism

Lobatchevski is not the only one to cast doubt upon Kantian epistemology. Mathematicians as eminent as Poincaré are likewise touched by a spirit of skepticism :

> *"The axioms of geometry are therefore neither synthetic a priori judgements nor experimental facts. They are conventions; our choice, among all possible conventions, is guided by experimental facts; but it remains free and is constrained only by the necessity to avoid contradiction. Thus, the postulates can remain rigorously true even when the experimental laws that led to their adoption are only approximate*[51]*."*

This position expresses a kind of moderate conventionalism, or nominalism. In Poincaré's eyes, no geometry can ontologically or epistemologically prevail over the others. Existence in mathematics is nothing other than non-contradiction, such that one may legitimately say of all these geometries that they exist. As for knowing which geometry best describes our physical world? Given the current state of our knowledge, this question is most often undecidable.

But does this mean that all geometries are equally valid? Quite the contrary. On a practical level, Poincaré puts forward two arguments in favor of Euclidean geometry[52]:

- **Simplicity**: On the one hand, it is the simplest representation, technically speaking. The formulas of spherical trigonometry are incomparably shorter and more elegant in this framework. True, there may exist certain integrals for which Lobachevsky's representation is more convenient, but for the vast majority of calculations and models a mathematician is led to employ, it remains the most effective. Poincaré clarifies: "and it is not such merely as a result of our mental habits or of some vague direct intuition we might have of Euclidean space; it is the simplest in itself, just as a first-degree polynomial is simpler than a second-degree polynomial[53]."

- **Agreement with experience**: On the scale of material objects familiar to us, Euclidean geometry aligns well with the physical world

[50] *Ibidem*, p. 210. The translation is mine.
[51] Poincaré [152], *op. cit.*, p. 67.
[52] *Ibidem*, p. 30.
[53] *Ibidem*.

and allows us to account for the properties of objects, their displacements, and so forth. Poincaré admits that matters might be altogether different if bodies were not, for the most part, solid, but rather liquid or gaseous, or if they underwent significant deformation during their movements. In such a case, it might have turned out that other geometries were more appropriate.

4.4 Kant Turns in His Grave

As suggested by the views of Poincaré and Lobachevsky, the new geometries deal a severe blow to the idea that there exist *a priori* structures of mathematical thought. In what precise sense are Kant's arguments thereby endangered? What becomes of synthetic *a priori* judgements? To address these questions, it may be helpful first to recall Kant's position. In his view, as stated in Chapter 2.3, space represents neither pure ideality nor pure reality. One must reject both the stance that regards it as an *in-itself* endowed with intrinsic properties, and the stance that sees it as a mere construction of the mind. I shall analyse the arguments he offers against these two theses and discuss them within the framework of non-Euclidean geometries.

4.4.1 Is Space a Thing in Itself?

Kant's first argument on this question—arguably the most important—consists in claiming that space is that through which things can be thought as outside of one another. Indeed, it is by means of space that one can conceive the fact that two things, whether identical or not in terms of their internal properties, may be situated in different places. Space, therefore, cannot be the product of an empirical operation. It is the very condition under which objects can be represented. As such, this argument stands independently of the geometry employed and is thus not undermined by the development of non-Euclidean geometries. Some might suggest that the fact of having an intuition of space does not preclude space from also being a property of the world. I shall leave that debate aside for now, in order to focus on a second argument.

In the *Prolegomena to Any Future Metaphysics*, Kant draws our attention to the epistemological status of enantiomers[54]. These objects are mirror images of one another and occupy a position analogous to that of my right hand with respect to my left hand. They differ in none of their internal properties, such that the complete description of one tells us everything about the other. And yet, we cannot say that a right-hand glove is identical to a left-hand glove. According to Kant, this example reveals a difference that is neither purely *internal* to the objects, nor purely *external* (or relational).

[54] Kant [101], *Prolegomena to Any Future Metaphysics*, § XIII.

Properties such as "being to the left of this" or "being above that" are, after all, merely relational. It is enough to alter the arrangement of objects to reverse these relations: what is on the right can be moved to the left, what is below can be moved above, and so forth. But in the case of gloves, no symmetry allows a right-hand glove to be worn on the left hand. It is a property that is, in a sense, "intrinsic," yet one that appears only when the object is considered in reference to the totality of space. This property is called chirality.

What does this mean for Kant? That the objects of our sensible experience are not things in themselves, for otherwise one could describe them solely in terms of internal or external properties. The thing in itself exists independently of the way it appears to us. Yet here, it would seem that what we encounter are merely manifestations of things in themselves, already structured by the *a priori* form of our sensibility. It is because objects are given in an *a priori* intuition that chirality appears. This property is the sign that objects can only be known within a space, and that this space is itself nothing but the mode through which we receive sensible objects :

> "We cannot render intelligible by any particular concept the difference between similar and equal, yet asymmetrical (incongruent) things; we can do so only by referring to the right and left hand, which involves an immediate relation to intuition[55]."

The emergence of non-Euclidean geometries puts this argument in jeopardy. In certain spaces, it is indeed possible to create symmetries that reverse chirality. A German philosopher by the name of Wittgenstein rightly writes: "The glove of the right hand could be put on the left hand if it could be turned round in four-dimensional space[56]." In fact, a similar phenomenon already appears in one dimension. The two segments a and b shown in Figure 4.5 are enantiomers in one-dimensional space. One indeed needs a second dimension to make them coincide.

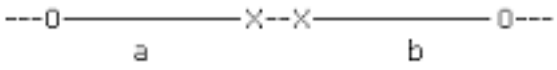


Figure 4.5: Wittgenstein's Line

Wittgenstein's objection can thus be summarized as follows: if space were non-Euclidean, it might be that enantiomers are in fact perfectly identical in themselves, and that their apparent difference stems only from our

[55] *Ibidem.*
[56] Wittgenstein, *Tractatus* [202], 6.36111.

limited spatial intuition. It would then be no more than a phenomenal illusion. That said, this kind of strategy still seems to play in Kant's favor. For he never claimed to have knowledge of noumena (or things in themselves). The mere fact of admitting that enantiomers can appear in our spatial intuition seems to attest to the existence of such an intuition. It thus remains possible to affirm that objects are given to us within a framework shaped by an *a priori* form of sensibility which alters their appearances. If Kant must concede some ground in admitting that space in itself might coexist with our intuition, he nonetheless retains the thesis that there exists an *a priori* form of sensibility—one capable of grounding a geometry distinct from that space itself.

4.4.2 Is Space a Construction?

In a spirit somewhat akin to Poincaré's, one might also ask whether our intuition—far from being an *a priori* structure of sensibility—is not, instead, a cultural or empirical construction. In our own world, bodies seem to be quite adequately described using Euclidean geometry, which has quietly imposed itself upon our mathematical descriptions of reality. But had we grown up in a hyperbolic world, is it not reasonable to think that our intuition of space would have been different? Kant's main response to this kind of skepticism is the following: to reject the idea that space is an *a priori* intuition (or that anything is *a priori*, for that matter) is to give up the possibility of explaining the existence of *a priori* mathematical propositions. In doing so, one renounces the grounding of universal truths. For if space derives from experience—whether sensory or linguistic convention—this means that everything we know about it is known only *a posteriori*. We are thus compelled to explain where the power of mathematical reasoning comes from, and what grants it greater stability and certainty—at least in appearance—than the other empirical sciences. Both positions are defensible, depending on whether one places more importance on the universality of mathematics or the relativity of our cultural representations.

It is worth noting that, for Kant, space is not an *a priori* concept either. A concept can certainly refer to an infinity of possible things—just as the concept of triangle applies to an infinite number of possible triangles—but no concept can itself be infinitely divisible. The concept of a straight line, for example, is not infinitely divisible: if one divides a straight line, one obtains two half-lines, and no longer a line. The notion of space in which the concept of the line is immersed, however, *is* infinitely divisible. This is why Kant cannot consider space a concept of the understanding, in the way idealists might, and even less a construction produced by the mind from sensory experiences. On the contrary, it is an *a priori* intuition that structures our concepts. It is, in other words, a condition of possibility for thought itself. Non-Euclidean geometries do not seem to undermine this

argument in any significant way. Discrete geometries, however, might. But delving into the details of that issue would lead us too far from our central concern.

4.4.3 Are Mathematics Synthetic *a priori*?

Perhaps the most serious objection is the following. Kant argues that all propositions in pure mathematics are synthetic *a priori* propositions. Furthermore, he holds that pure intuition of space and time is the *a priori* form that renders these judgements possible. If our intuition is, as it appears to be, limited to Euclidean geometry, *how can one explain that valid mathematical propositions have been conceived concerning non-Euclidean geometries*? One must either (a) conclude that these propositions are not synthetic *a priori*, or (b) maintain that our intuition of space somehow encompasses non-Euclidean geometries.

In case (a), what would become of the status of mathematical propositions? One option is to claim that elliptic and hyperbolic geometries arise from analytic judgements, while Euclidean geometry remains synthetic *a priori*. The problem then becomes justifying this difference. Are there truly any remarkable distinctions in their methods of reasoning? In both cases, the method is hypothetico-deductive in nature. The same rules of inference, the same symbols, and—supreme irony!—nearly the same initial hypotheses guide thought in either case. The only real difference lies in a single postulate.

But perhaps it will be argued that the relevant difference lies in the content. The propositions of Euclidean geometry, it may be said, correspond to *a priori* intuitions, whereas those of non-Euclidean geometries have lost contact with sensible content. This observation is entirely correct. But if one takes Kant's doctrine seriously, one must conclude with him that the assertions of non-Euclidean geometries, in extending beyond intuition, are no more than empty speculations. Their fate should be similar to that of the antinomies of reason: sooner or later, contradictions should arise. Yet it has been demonstrated that these theories are true—or, more precisely, that they can be no more false than Euclidean geometry itself. By interpreting each term of a non-Euclidean geometry through the terms of Euclidean geometry, it is indeed possible to show that if the latter is consistent—as was proven in 1951 by Tarski—then the former is as well. Moreover, the progress made in the field over the past one hundred and fifty years suggests that, unlike speculative metaphysics, non-Euclidean geometries form a body of stable knowledge which, far from collapsing over the decades, has only grown and flourished into a prosperous domain.

An alternative would be to view the propositions of non-Euclidean geometries as synthetic *a posteriori* judgements, while preserving the synthetic *a priori* status of Euclidean geometry. But this position seems even more

difficult to maintain. It would imply that the results of elliptic geometry—which are notoriously hard to visualize—are the fruit of experiments carried out in our physical space. Yet this space, at the scale of our direct observation, is far closer to Euclidean geometry than to any other form of geometry. How, then, can a Euclidean space serve as the basis for the experience of a non-Euclidean one? This, without even mentioning the contradiction it presents with the history of science. For nearly two millennia, mathematicians consistently rejected non-Euclidean geometries. It was only through the development of logical analysis that these geometries gradually came to be accepted by the scientific community as viable possibilities. Only after this theoretical breakthrough did physics—borrowing its tools, as it often does, from the queen of the sciences—translate these new geometries into cosmology. Our empirical reality, far from having contributed to their development, seems on the contrary to have been their principal obstacle, at least in the beginning. And as if all these arguments were not enough, we are still left to ask—as we did in the previous paragraph—what methodological or content-based difference between Euclidean and non-Euclidean geometries could possibly justify an epistemological distinction between them?

Since it appears difficult to assign to the propositions of Euclidean geometry a status different from that granted to non-Euclidean geometries, the only remaining possibility is to revise the belief that *all mathematical propositions are synthetic a priori judgements*. This opens up multiple alternatives. First, one may suppose that they are, in the end, merely analytic judgements — and that Kant was mistaken in wanting to move beyond this. This is the path taken by the logicists when they attempt to derive mathematics from an appropriate formal system. Another solution would be to see them as synthetic *a posteriori* judgements. This leads to the philosophical positions of Poincaré (in support of conventionalism) and Lobachevsky (in favor of empirical constructivism). The price to pay, as we have already noted, is a loss of universality. It becomes difficult to defend the aprioricity of mathematics — at least not in the pure and rigorous sense Kant had in mind[57].

Case (b) also seems to represent a viable path. It consists in substituting a new form of pure intuition for that of space and time. This intuition can no longer take the form of a simple self-evidence, but must accommodate the fact that, in the case of Euclidean geometries, objects are given to us only through intuitive interpretations. Hyperbolic geometry, for instance, can be conceived by way of the Poincaré disk; elliptic geometry can be visualized on a sphere in which antipodal points are identified. All these simplified models are projections onto a plane accessible to human intelligence. Intuition is

[57] Empiricism must also account for the fact that we possess the faculty to structure the data of our experience. How can it be that this capacity, which renders experience possible, might itself come from experience?

therefore no longer that horizon of clear and distinct givenness, where things appeared identical to themselves. It is, at best, a space where objects are given to be seen from a particular angle. This idea is not without echoing several philosophical conceptions from the 20$^{\text{th}}$ century. I am thinking, on the one hand, of Bohr's principle of complementarity, and on the other, of Husserl's *Abschattungen*.

Let us note, in concluding this discussion, that although the pure intuition of time enjoyed a more comfortable position than that of space throughout the 19$^{\text{th}}$ century, it was not spared upon entering the 20$^{\text{th}}$. Advances in physics showed that time is not independent from the frame of reference. In 1905, Einstein published a now-famous article demonstrating that time dilates as a function of the velocity of the Galilean reference frame, in accordance with the (already known) Lorentz transformations. Ten years later, he constructed a model capable of generalizing these results to all reference frames. These important developments, carried out with the collaboration of Minkowski, provided the occasion to introduce a four-dimensional representation of space-time, which ultimately disrupted (1) our intuition of time and (2) the belief in an intuition of time independent from that of space. That said, these advances do not mark the death of intuitionism as such — quite the contrary! Kant would continue to influence entire generations of mathematicians, foremost among them Hilbert, whom we shall soon see offering a new epistemological conception of axiomatics.

What to take away from this chapter

Kant presupposed the existence of two innate structures in the human being: *intuition of time* and *intuition of space*. It is convenient to use these intuitions to explain the differences in judgement between the human mind and our contemporary algorithms.

However, the history of non-Euclidean geometries seems to contradict the idea that these intuitions are as necessary as they appear to be for the exercise of judgement. One might wonder, in particular, whether a *synthentic a priori judgement* truly exists.

This is why it seems appropriate to begin a search for the most fundamental structures of the human mind. If it is not the intuition of time and space, what could it be?

I will attempt to answer this question in the next part of this dissertation. But first, we must understand how the abandonment of Kant's theory gave rise to the theoretical development of the first computers. That is the subject of the next section.

Chapter 5

Digital Rationality: The Birth of the Computer

> By "calculation" I mean here any solution that represents reasoning, even if it has nothing to do with numbers.
>
> Gottfried Wilhelm Leibniz

The upheaval triggered by non-Euclidean geometries stems less from the abandonment of Euclid's fifth postulate than from a more startling realization: mathematical truths can overflow the bounds of our intuitions. Since Antiquity, it had already been known that the shortest path between two points on the Earth's surface is a geodesic. It was also known that the sum of the angles of a triangle drawn on a sphere exceeds 180 degrees. Such results did not offend our habitual experience of space, nor our accepted representations.

But in the 19^{th} century, it was suddenly discovered that pure reasoning holds an unsuspected power. By changing the premises of Euclid's well-established theory, and by exploring the logical consequences without invoking human intuition, entirely new worlds—strange and hallucinatory—unveiled themselves before the human gaze. A conviction gradually imposed itself: formalism surpasses intuitive content. Truth, far from remaining quietly within the realm of synthetic *a priori* judgements, falls under the dominion of analytic judgement. It flows entirely from logical deductions derived from a small number of axioms. These axioms can be chosen in different ways, depending on pragmatic needs and established conventions, without relying on sensory perception, intuition, or the understanding of the mathematician.

Such considerations are troubling for the philosophical project I have defended thus far. My thesis held that the synthetic *a priori* judgement

might well be the one element that still eludes our algorithms. This claim seemed to be under the best of auspices. It appeared to be supported by the philosophical theory of judgement (from Plato to Kant) and by algorithmic practice (the ARC tasks). But a shadow has come to darken the hope of an easy triumph: in reviewing the history of non-Euclidean geometries, it became apparent that Kantian notions of pure intuition of space and time might be at fault.

Does this mean that synthetic *a priori* judgement must be abandoned? Should it be seen as a remnant of human thought, akin to the caloric fluid, the ether, the boreal fluid, animal magnetism, and all those defunct scientific theories? That would be somewhat radical. It still seems correct to say that our representation of the world fits within a Euclidean geometry, and that our representation of time follows the Newtonian timeline. One must also grant Kant the visionary merit of having defended the idea that these intuitions are within us, that they are structures of representation, and that they do not necessarily correspond to the true reality of the world.

Despite the heavy clouds now hanging over the notion of synthetic *a priori* judgement, it still seems promising to think that there is indeed a structure, within the human mind, that leads us to perceive information differently than a machine does. Have I not shown, through optical illusions, that humans represent data holistically? Have I not suggested that the machine struggles to perform abstraction? That is why the essential task before me, if I am not to sign the death warrant of my endeavour, will be to offer a new definition of synthetic *a priori* judgement. This definition must avoid the two pitfalls of (1) presupposing the existence of a Euclidean intuition of space-time, and (2) reducing thought processes to mere logical calculations.

This thorough revision of the concept of synthetic *a priori* judgement will first require a more precise understanding of the paradigms of theoretical computer science that gave rise to the dream that "everything is computable". By identifying the limitations of these paradigms and understanding what kinds of questions they fail to answer satisfactorily, I hope to gather the ingredients necessary to construct a new theory of human judgement. This strategy will also allow for a deeper understanding of the functioning of the Machine—understood as a specific mode of relation to information.

In this chapter, I would like to resume the study of human judgement at the point where non-Euclidean theories left off: with the triumph of pure reasoning over spatio-temporal intuitions. From this barren and desolate ground, I aim to show that the progressive abandonment of Kant's *a priori* forms led philosophy to represent the world solely through computable propositions. To every scientific question—precise and rigorous—there must correspond an equally exact answer. The process that leads from question to answer is governed by clear, automatable rules, similar to Euclidean di-

vision. If you wish to compute the division of ten by three, the procedure to follow is predetermined. You have nothing to do but follow the steps. Likewise, the hope for science is that it might be automated in such a way as to leave no room for error. This dream, this ideal, was popularized by the mathematician David Hilbert. It gave birth to the very first form of artificial intelligence: a human being equipped with a pad of paper and a pencil, applying rules of calculation, deriving equations, producing symbols on the blank page—without ever invoking their own judgement.

5.1 The Dream of a Universal Calculator

On August 6, 1900, the Palais des Congrès is buzzing with excitement. Scholars in bowler hats, briefcases under their arms and pipes in their pockets, parade beneath the arcades. Some are talking geometry, others algebra and arithmetic. One among them, whose hat is more imposing than the others, takes the initiative to open the session. David Hilbert, together with Henri Poincaré, introduces the International Congress of Mathematicians. During his keynote address, which will later take place at the Sorbonne, he lists twenty-three open problems in mathematics.

One of them, the second problem, is of particular interest to us. It raises the question of whether arithmetic can be axiomatized in a complete and consistent way. Arithmetic refers to the set of positive integers $0, 1, 2, \ldots$, along with the operations of addition and subtraction. This arithmetic is said to be complete if it is possible to provide a finite list of axioms (that is, propositions assumed to be true in themselves) from which all arithmetical truths can be derived, such as "5 + 7 = 12" or "every even number can be associated with one and only one odd number". Moreover, this arithmetic is said to be consistent if no contradiction can result from this list of axioms.

To illustrate the problem, it may be useful to note that Euclidean geometry possesses a complete and consistent axiomatic system. All the truths that can be stated about lines, points, triangles, and circles in the plane can indeed be proven from a small set of axioms known as Euclid's axioms:

- **Axiom 1:** There always exists a straight line passing through any two points in the plane.
- **Axiom 2:** Every segment can be extended into an infinite straight line.
- **Axiom 3:** From a given segment, there exists a circle whose centre is one of the segment's endpoints and whose radius is equal to the segment's length.
- **Axiom 4:** All right angles are equal to one another.

- **Axiom 5:** Given a point and a line not passing through it, there exists exactly one line through the point that is parallel to the first.

Now it so happens that, in Hilbert's time, no one had yet managed to produce an analogous list of axioms for arithmetic. Hilbert thus imagines the fiction of a mathematician equipped with a pencil and an unlimited supply of paper. Could such a person, using only a finite combination of symbols, write a proof for every true arithmetical statement, and only for those?

Rather than insisting on the existence of abstract entities such as numbers, operations, or algebraic groups, Hilbert considers the *sign* as the only acceptable definition of a mathematical object. This sign is utterly tangible: it exists on a sheet of paper, obeys strict syntactic rules, and all the operations it can undergo are visible, traceable, verifiable. The problem of axiomatizing arithmetic thus becomes the problem of finding the complete set of rules—an algorithm—that governs arithmetic. In particular, it must be ensured that:

- The algorithm consists of a finite set of simple instructions, each described using a limited number of signs;
- The algorithm produces a result in a finite number of steps;
- The algorithm requires no more intelligence, on the part of the executor, than what is needed to follow the instructions.

We might say that Hilbert lays the groundwork for the concept of computability, as it would later be formalized by mathematicians Church and Turing in their eponymous thesis (the Church–Turing thesis). It is not by chance, after all, that the English word *computer* once referred to a human being capable of performing calculations on paper—someone tasked with computing resistance coefficients, tensions, or ballistic trajectories. It was only after the development of the first machines that the term would gradually shift to mean what we understand today.

Hilbert's speech would be followed, within a few decades, by two types of results showing—in two distinct frameworks—that it is impossible to express all mathematical truths using rules, a pencil, and a sheet of paper. The two frameworks in question are, on the one hand, theories of provability, and on the other, theories of computability. The former represent reasoning via formal axiomatic systems in the style of Euclid, while the latter define reasoning as algorithmic, in the style of Hilbert.

Définition 2 (Paradigm of Provability)**.** *Every computation can be expressed as a logical deduction within a formal system. A formal system is composed of:*

- ***D'axiomes :*** *Propositions assumed to be true in themselves.*
- ***De règles de raisonnement :*** *Laws determining how one proposition can be derived from another.*

Définition 3 (Paradigm of Computability)**.** *Every computation can be performed by a Turing machine. The latter consists of:*

- ***A tape:*** *A strip of paper divided into an infinite sequence of cells. Each cell may contain a symbol or remain blank.*
- ***A read/write head:*** *A mechanical device capable of moving along the tape, reading symbols, erasing them, and writing new ones.*
- ***A program:*** *A set of instructions telling the read/write head what to do. For example: "If the symbol 'a' appears in the first cell, erase the symbol in the third cell."*

A "computation" consists in posing a question to the machine by writing it on the tape, and reading the answer it produces—after a finite amount of time—on that same tape.

These two paradigms echo, in a way, Leibniz's dream of a universal characteristic, when he exclaimed, regarding any philosophical or theological problem: "Let us calculate!" Yet two major mathematical results show that, for each of these paradigms, there exist certain uncomputable problems. One might even say that the only problems solvable through calculation are mere isolated points in the infinite space of possible mathematical propositions. But due to the limitations of the human mind, only the most evident solvable questions have received sustained attention in the history of mathematics—so much so that the submerged part of the iceberg was long mistaken for nonexistence.

Here is a foretaste of the two results in question. I will take them up one by one and elaborate on them in the coming chapters.

Theorem 1 (Gödel's Incompleteness Theorems)**.** *Every axiom system A ssufficiently strong to express the propositions of arithmetic contains at least one undecidable proposition (whose negation is also undecidable). This is particularly the case for the proposition stating: 'A is consistent '.*

Theorem 2 (The Halting Problem)**.** *It is impossible to construct a Turing machine capable of determining, for every other Turing machine, whether it will provide an answer to a given problem in a finite amount of time—or whether it will run indefinitely without ever stopping.*

5.2 Theories of Provability

Gottlob Frege is, in every possible way, a caricature of the logician. He is precise, austere, obstinate—and radically devoted to intellectual integrity. His nature, with that kind of inevitability known only to nature itself, drew him toward the project of uncovering the axioms of arithmetic. This pursuit enchanted him early in his career. He imagined producing, for addition, subtraction, and multiplication of whole numbers, what Euclid had once done—millennia before him—for squares, circles, and triangles. Such a feat would offer a concrete, operational answer to David Hilbert's dream of automating mathematical thought.

At the root of this philosophical project, Frege begins by redefining the concepts of the "synthetic" and the "analytic". To determine the nature of an analytic proposition, Frege claims, one must analyze the steps of its demonstration, beginning at the conclusion and tracing backward to the axioms. If, at the end of this examination, it turns out that the proposition depends solely on logical laws, then it is said to be *analytic*. But if even a single law drawn from other sciences is involved, then the proposition is *synthetic*.

Moreover, a proposition is said to be *a priori* if it relies only on fundamental laws—such as the principles of Newtonian mechanics—and *a posteriori* if it incorporates empirical data. According to this definition, all analytic propositions are thus *a priori*. But they are far from mere tautologies. Frege's position, his fundamental thesis, consists in claiming—contra Kant—that all arithmetic propositions, such as "7 + 5 = 12", are analytic[1]. This stance lays the groundwork for logicism. It opposes both Kant, for whom mathematical propositions are synthetic *a priori*, and English-speaking thinkers like Henri Poincaré or John Stuart Mill, for whom they are synthetic *a posteriori*.

Frege sets out to prove this logicist thesis in his magnum opus, entitled *The Foundations of Arithmetic*. To this end, he begins by defining a set of basic laws and rules of inference, along with definitions that assign denotations to these concepts. These definitions do not, in themselves, create anything new; they merely name the terms at play. As he puts it himself: "The definitions do not really create anything, and in my opinion may not do so; they merely introduce abbreviated notations (names), which could be

[1] Bélanger [114], *La méthode axiomatique durant la crise des fondements* in *La crise des fondements : quelle crise ?*, p. 46.

dispensed with were it not that lengthiness would then make for insuperable external difficulties[2]." And as for the status of the basic laws? Their role is foundational. They are axioms "evident in themselves."

What precisely is meant by this "self-evidence", the cornerstone of Frege's thinking, is the subject of an entire strand of philosophical literature. According to Robin Jeshion, from the University of Southern California, the term carries two distinct meanings:

> More specifically, my view is that Frege had two notions of self-evidence. One notion is that of a truth being foundationally secure, yet not grounded on any another truth. The second is that of a truth that requires only clearly grasping its content for rational, *a priori* justified recognition of its truth[3].

The basic law should be "self-evident" in both senses of the term before it can be accepted. The first sense appears to express a pragmatic requirement: that the axiom must not depend on any other axiom. The second seems to invoke an epistemic criterion—clarity, in a near-Cartesian sense. Frege does not, therefore, presuppose the existence of an intuition in which such self-evidences might be made manifest. The kind of evidence he has in view is rather that of a linguistic act.

Be that as it may, Frege, armed with this self-evidence, sets out in search of the axioms that underpin arithmetic. After several years of inquiry, he records his conclusions in a book destined for a tragic fate.

5.2.1 The Barber Paradox

Frege's *Foundations of Arithmetic* was published in 1884, but the project it carried was unexpectedly cut short when Russell identified the paradox that now bears his name. In the letter he sent to Frege in 1902, Russell explained that the way mathematical objects are defined seems to lead to contradictions. He pointed in particular to the concept of a "class":

> *Likewise there is no class (as a totality) of those classes which, each taken as a totality, do not belong to themselves. From this I conclude that under certain circumstances a definable collection [Menge] does not form a totality*[4].

Frege would write back six days later: "It seems that my Rule V (§20, p.36) is incorrect, and that my explanations in §31 are insufficient to ensure that my combinations of signs have meaning in all cases[5]."

[2] Frege, *The Foundations of Arithmetic*, p. vi.

[3] See Jeshion [91], *Frege's Notions of Self-Evidence*.

[4] Letter from Russell, June 16, 1902. Quoted from Linsky, *Ernst Schroeder and Zermelo's Anticipation of Russell's Paradox* in *La crise des fondements : quelle crise ?*, pp. 8–9.

[5] Frege's response to Russell, June 22, 1902. Quoted from *ibidem*, p. 8.

It is possible to give this Russell paradox a more intuitive illustration, known as the *Barber Paradox*. Let us imagine, for this purpose, a barber whose shop is daily plagued by endless queues. To reduce the waiting time, the town council passes the following resolution: "Henceforth, the barber must shave all those who do not shave themselves, and no one else." Though reasonable at first glance, this new law will have, for our good man, an unforeseen consequence. On the seventh day, a logician by the name of Russell arrives in the village. Pipe in mouth and hat under arm, the newcomer asks, in true Sherlock Holmes fashion: "Tell me, dear barber, under the laws of your village, can you shave yourself?"

The question is thorny. For if the barber answers yes, he breaks the law, since he is forbidden from shaving those who shave themselves. But if he answers no, he still breaks the law, since he is supposed to shave all those who, like him, do not shave themselves. History does not tell us what became of the barber, but it is said that before the court, he pled guilty[6].

A second, even more succinct, illustration of the paradox comes from Martin Gardner[7]. It consists of imagining an encyclopaedia that lists the names of all encyclopaedias that do not list themselves. The paradox arises when one poses the following question: must Gardner's encyclopaedia list itself?

What may seem like an innocent paradox plunges logicism into a state of deep perplexity. The only way for Frege to circumvent the antinomy is to revise Rule V so as to prevent any class from containing itself. More broadly, every logicist writing after this 1902 letter must reckon with this difficulty. As Poincaré points out in *Mathematics and Logic*, the vicious circle arises from assuming that a collection of objects can contain elements that, in turn, can only be defined by means of the collection taken as a whole[8].

For example, the proposition P, which asserts that "All propositions are either true or false," is problematic. Indeed, to which propositions does P refer? Doubtless, one would say it concerns "all existing propositions." Yet how can we define the set of "all propositions" referenced in P, if the mere act of stating P already creates a new proposition? It is therefore necessary to abandon terms such as "all propositions" and, more generally, terms that refer to all members of a set[9].

---

[6] A variant tells that he let his beard grow until it was long enough to strangle all the logicians in the village...

[7] At least, it is attributed to him, though I have been unable to find the exact reference.

[8] Poincaré, *Les mathématiques et la logique*, p. 307.

[9] Terms referring to all members of a set are not always problematic—in fact, they are only rarely so. For example, one may assert: "All natural numbers are greater than or equal to zero," and there is nothing paradoxical about this. However, a logic permitting quantification over all members of an infinite set can give rise to paradoxes, such as Russell's paradox. One way to avoid such paradoxes is to impose additional constraints on general terms, a process known, in technical terms, as infinite quantification.

5.2.2 Cantor's Paradox

Another paradox of major importance in the history of mathematics is the one demonstrated by Cantor in 1899. Although it precedes Russell's paradox chronologically, it is more convenient, from a strictly logical perspective, to present it second. Rather than focusing on the set of all sets that do not contain themselves, this paradox considers the set of all sets—period. In both cases, we are dealing with a non-constructive procedure: we posit the existence of an infinite set without explicitly specifying its contents. However, while Russell's paradox involves an evident contradiction, it is not immediately clear what logical issue arises from speaking of the set of all sets.

I will attempt to explain the paradox. As a first observation, it is important to note that the set of all sets, denoted ENS, is infinite and self-referential. Infinite, because there are infinitely many possible sets. Self-referential, because by definition, E must include all sets—including itself. Thus, there exists an element $x \in E$ such that $x = ENS$. This means that the set ENS can be imagined as a great bag that contains, within it, a bag that is itself. One senses that something about the notion of membership here clashes with common sense. How can something contain itself? And yet, it is not uncommon to encounter counterintuitive properties when dealing with infinite objects.

To rigorously demonstrate the existence of a paradox, Cantor uses an argument known as the "diagonal argument." I will show how this argument can be used to demonstrate that, in a universe containing only natural numbers $1, 2, 3, \ldots$, the set of all sets of natural numbers cannot contain itself. I will then generalize the reasoning to the set ENS of all sets. Readers wishing for more detail are invited to consult Appendix 4, which presents Cantor's theory in full depth and breadth.

Cantor's Diagonal Argument

The diagonal argument can be used to show that the set $\mathbb{N}$ of natural numbers is not large enough to contain the set of subsets of natural numbers, denoted $\mathbb{P}(\mathbb{N})$. To recall, a subset of the natural numbers is any set that includes zero, one, or several natural numbers. For example, $1, 5, 7$ is a subset of $\mathbb{N}$, as is the set of even numbers $2, 4, 6, \ldots$, the empty set , and $\mathbb{N}$ itself.

To reach this result, let us suppose that there exists a book L with a countably infinite number of pages, and that each page lists a distinct set of natural numbers. Page number 1, for instance, might contain the set $1, 2, 3$, page 2 the set 7, and so on. Cantor shows—within a formalism uniquely his own—that if one admits the existence of such a book, a contradiction ensues. This proof strategy, which consists in assuming the existence of what one seeks to refute in order to reveal its absurdity, is known as *reductio*

ad absurdum. We have already encountered this kind of strategy in our discussion of non-Euclidean geometries (see Section **??**).

Let us therefore assume that Cantor's book exists. Let us call an "ordinary number" any natural number n that is not listed on the page numbered n in the book. In our example, the number 2 is an ordinary number, since the set represented on page 2, namely 7, does not include the number 2. The number 1, on the other hand, is not an ordinary number in our example, since page 1 contains the number 1 in the set it displays: $1, 2, 3$.

Since Cantor's book includes all sets of natural numbers, there must exist a page—let us say page n'—that contains the set of all ordinary numbers, and no others. The set represented on page n' must thus include, among others, the number 2, but not the number 1. Are you still with me? The question we now ask is the following: is n' itself an ordinary number?

Suppose first that n' is an ordinary number. Then it must be listed on the page that contains all the ordinary numbers. But that page is page n'—which contradicts the definition of an ordinary number. Suppose now that n' is not an ordinary number. By definition, it must then appear on page n'. However, this page includes only the ordinary numbers, and nothing else. We conclude that Cantor's book is paradoxical.

If we wish to prevent mathematics from contradicting itself, we must reject the existence of Cantor's book. What does this imply, logically speaking? That it is impossible to enumerate all subsets of natural numbers by listing them in an infinite book. In other words, it is impossible for the book that contains all natural numbers to also contain the set of all subsets of natural numbers.

The Set of All Sets

Cantor's diagonal argument has shown that the set of natural numbers $\mathbb{N}$ is not large enough to contain the subsets of natural numbers. But the argument can be generalized to any kind of set. For instance, one can use it to show that the set of real numbers is not large enough to contain within itself all subsets of real numbers, denoted $\mathbb{P}(\mathbb{R})$. This latter set, in turn, is not large enough to contain the set $\mathbb{P}(\mathbb{P}(\mathbb{R}))$ of subsets of $\mathbb{P}(\mathbb{R})$, and so on. There thus exists a growing hierarchy of sets, each nested within the next. The question we now face is whether there exists a kind of super-set, ENS, that could contain all possible sets—including itself.

The answer is no. And with good reason: Cantor's diagonal argument can be applied to ENS itself. If we suppose such a set exists, then it would not be large enough to contain the set $\mathbb{P}(ENS)$ of its own subsets. In other words, it is always possible to imagine a set larger than the largest set one can conceive. It follows that the set of all sets is ultimately just as contradictory as the set of all sets that do not contain themselves—as imagined by Russell.

5.2.3 Circumventing the Paradoxes

Various attempts are being made in parallel to circumvent Cantor's paradox and the barber paradox. One initial solution consists in assigning each term in the theory a certain type. The minimal type corresponds to individuals; next, there is the type of sets or predicates of individuals, then the type of sets of sets or predicates of predicates, and so on. With this solution, a set cannot contain itself, but it can only contain objects associated with the lower type. The theory of types also stipulates that there are properties which can apply to themselves, and others which cannot, in order to avoid generating paradoxes. This theory of types was inspired by Russell and Whitehead's *Principia Mathematica*. The latter, however, presented not only types but also a hierarchy of orders, which made it difficult to handle.

Another approach, which has met with the greatest historical success, is the Zermelo-Fraenkel formalisation, denoted ZF. This theory describes what a mathematical set is, potentially of "infinite" size[10]. Although it is not, strictly speaking, a theory of mathematical logic, it is remarkable in that it can be expressed within first-order logic. Moreover, it adequately avoids the paradoxes that arise in the foundational approach to mathematics by employing a set of fairly intuitive axioms. One could say, in a certain sense, that the crisis of foundations has been resolved. However, to do so would be to overlook not only the technical nature of that crisis, but also the more fundamental epistemological question that it raises: what is the status of the axioms? Are they sufficient to support *all* mathematical results? And are they, moreover, sufficiently *self-evident*? In what sense?

A number of axioms, such as the Axiom of Choice (AC) and the Continuum Hypothesis (CH), are independent of the basic axioms of ZF theory and can be adjoined to it. The two new resulting theories are denoted ZFC and ZF + CH, respectively[11]. The addition of such axioms makes true certain theorems that defy intuition. On the other hand, omitting them renders the theory incapable of deriving some fundamental results in algebra and analysis.

I shall illustrate this situation with the Axiom of Choice. In its simplest form, this axiom states that "every set S which does not contain the empty

[10] The size of a finite set depends on the number of elements it contains and is called its cardinality. The size of an infinite set can be understood in several ways. Rigorous definitions require the introduction of transfinite numbers. See Appendix 5 for further details.

[11] The independence of the Axiom of Choice was shown by Cohen in 1969. By this time, it was already known, thanks to a result by Gödel dating to 1930, that if ZF is consistent, then ZF $\nvdash \neg$ AC. In 1969, Cohen succeeded in constructing a model of ZF in which the negation of the Axiom of Choice is valid, showing that if ZF is consistent, then ZF $\nvdash$ AC. These two independent results allow us to conclude that if ZF is consistent, then the Axiom of Choice is independent of the axioms of ZF. An analogous result is obtained with the Continuum Hypothesis: in fact, the hypothesis is valid in Gödel's model but not in Cohen's. For more details, see Van der Linden, *Notions de logique mathématique*, p. 50.

set has a choice function." This choice function associates, to every x in S, one of the elements of x. Formally, a function $f : S \to \cup S$ is a choice function if $\forall x \in S, f(x) \in x$. Intuitively, the Axiom of Choice therefore states that if one is presented with a collection of non-empty boxes of pralines, it is possible to define a function that will select in advance which praline to choose from each box as it appears before us.

At first glance, there is nothing in this axiom to alarm a mathematician. And yet, the subtlety lies in the fact that it remains applicable even when one is faced with a box containing an infinite number of pralines. In fact, a number of astonishing results can be shown on the basis of this axiom, including the following: by using the Axiom of Choice, a sphere of a given volume V can be decomposed into infinitely many parts, and then reassembled to form a sphere of arbitrary volume. The consequences of the axiom are, therefore, patently at odds with our intuition of space[12].

The Axiom of Choice is nevertheless indispensable for the demonstration of a great many theorems, beginning with the so-called Zermelo theorem, which states that every set can be well-ordered. This result is foundational in set theory and, indeed, was the objective pursued by Cantor in his early classification of sets. It guarantees (1) that there exists a minimal set (the empty set), and (2) that all sets can be classified according to their size. For any sets X and Y, whether finite or infinite, one can say whether $X < Y$, $X = Y$ or $X > Y$. The symbols $>, <$, and $=$ are then said to define an ordering relation.

The Axiom of Choice is also necessary for the proof of several results in algebra and analysis, including the following[13]:

- **Krull's Theorem**: every unital ring contains maximal ideals;
- **Tychonoff's Theorem**: the product of compact spaces is compact;
- **Nielsen–Schreier Theorem**: every subgroup of a free Abelian group is itself free Abelian;
- **The existence of non-measurable sets** (in the sense of Lebesgue);

[12] This illustration of the Axiom of Choice is not intended to retrace the genuine issues that the theory ZFC poses for mathematicians. The theory ZFC, in fact, contains many more counter-intuitive results than the decomposition of the sphere mentioned here. Indeed, the terms of ZF theory are highly abstract: they concern sets of entirely *arbitrary* objects, which do not directly refer to any tangible intuitive reality. This is why, in their practice, mathematicians tend not to be overly concerned with the intuitive interpretations of these concepts, but rather with their mathematical utility. It falls fundamentally to philosophers to enquire as to the significance of ZFC. To pursue this question rigorously, it would be essential to go beyond the textbook example of the Axiom of Choice, and to study conscientiously (1) the principal mathematical objects appearing in the theory, (2) the properties of these objects, and (3) the relations of implication between this theory and other mathematical frameworks.

[13] From Van der Linden, *Notions de logique mathématique*, p. 48.

- **Hahn–Banach Theorem**: the existence of a linear functional on a Banach space satisfying certain conditions—the values are prescribed on a part of the space, but are extended everywhere;
- **Stone's Theorem**: a duality between the category of Boolean algebras and that of compact totally disconnected Hausdorff spaces.

Finally, it is essential for establishing some elementary properties in $\mathbb{R}$, such as the equivalence between two common definitions of continuous functions: the epsilon-delta definition and the notion in terms of convergent sequences[14]. Given all this, can we, in all conscience, dispense with the Axiom of Choice? Or does the price to be paid for relinquishing it appear, instead, inadmissible? And if so, how then might we establish the mathematical properties listed above?

5.2.4 Gödel's Theorems

One might believe that the paradoxes of Russell and Cantor sound the death knell of the project to axiomatize mathematics. This is where another logician steps in. Our man is at least as precise, austere, tenacious, and upright as Gottlob Frege — but with something more tortured about him, more haunted. Kurt Gödel, of Austrian origin, is known for his heightened sensitivity. For his fear of invisible things: bacteria, atoms, and all that can awaken hypochondria. At Princeton, he is Einstein's closest friend. He will spend the final years of his life trying to prove the existence of the Devil. He is also one of the rare few to have read Russell and Whitehead's *Principia Mathematica*. In its entirety.

His reputation as a genius rests on two demonstrations concerning the incompleteness of arithmetic — that is, the impossibility of proving certain truths about the addition, subtraction, and multiplication of whole numbers. These demonstrations offer a response to Hilbert's problem regarding whether all true propositions can be proven by means of an automatic process consisting in deriving equations on a sheet of paper. Alas for the advocates of these authentic pencil-and-paper machines, the answer is negative. To prove it, Gödel drew inspiration from the famous Epimenides paradox. This paradox consists in considering a sentence P that says of itself:

> *Sentence P is false.*

If this sentence is false, it must be true. But if it is true, then it must be false. This paradox has been known since Antiquity. Ancient treatises on logic are riddled with liars who claim of themselves that they always lie. These linguistic games, despite the curiosity they invariably provoke,

[14] *Ibidem*. The first definition implies the second, but the converse does not hold in ZF without the Axiom of Choice.

were not considered worthy of philosophical attention. Gödel's great insight was to take inspiration from this paradoxical sentence P in order to show that there exist arithmetic truths which are indeed true, but rigorously unprovable.

Before Gödel's results, Hilbert's formalist school hoped (or preferred to believe) that every true proposition was provable, and that every false proposition's negation was, likewise, provable. Gödel, however, established that there exist true but unprovable propositions, thus beginning a complex and nuanced divorce between semantics (truth and falsity) and syntax (provability).

For Gödel's idea to work, the axiom system A, which allows one to construct the sentence P, must satisfy at least two conditions: first, (1) this system must exist, and (2) it must be mathematically interesting. Condition (1) means that one must be able to find axioms which allow the construction of a sentence that refers to its own provability. This is far from trivial. In an axiomatic system such as Peano arithmetic, the objects manipulated are natural numbers, together with operations like addition and subtraction, etc. These objects do not directly reference the notion of provability. This notion is not, in fact, an intrinsic property of arithmetic, but rather a property of the system that represents arithmetic. Formally, one calls this a meta-property.

Condition (2) states that our axiom system A, in addition to referring to its own provability, must also be mathematically substantive. What I mean by this is that it must be sufficiently powerful to speak about mathematical objects such as those manipulated in arithmetic. Otherwise, Gödel's result could be dismissed as a mere "logical curiosity"—there would be scant interest in knowing that a certain system of axioms, which no one uses and which cannot express relevant mathematical truths such as Fermat's Theorem or Goldbach's Conjecture, possesses the peculiar property of having true yet undecidable propositions.

To resolve point (1), Gödel devised an axiom system capable of representing all logical formulae of the form "$0 = 0$" by means of unique whole numbers. If two numbers are distinct, then they denote two different formulae; conversely, if two logical formulae are distinct, they correspond to different numbers. In cryptography, this is known as encoding.

If the encoding is sufficiently well executed, it then becomes possible to reduce the provability of a sentence to an arithmetical calculation. A formula such as "$0 = 0$" will be proven true if there exists a calculation on the numbers leading to that conclusion. Likewise, a formula "$0 = 1$" will be proven false if a calculation can be performed to show its negation, i.e., "$0 \neq 1$."

Within this paradigm, it becomes possible to express the sentence P: "This sentence P is not provable." This P is now a number within arithmetic, and the question of whether P is provable becomes a matter of calculation.

If such a calculation can be produced, then P is provable. But since P says of itself that it is not provable, this leads to a contradiction. It follows that P is indeed unprovable, yet it is not possible to prove this.

It is important to note that this axiomatic system, produced by Gödel, is capable of performing calculations on the natural numbers. It is therefore sufficiently powerful to express various arithmetical propositions, such as Fermat's Theorem or the Goldbach Conjecture. Condition (2) is thus genuinely satisfied.

This strategy, adopted by Gödel, enabled him to prove the first incompleteness theorem that now bears his name:

Theorem 3 (Théorèmes d'incomplétude de Gödel)**.** *Let A be an axiom system sufficiently strong to express the propositions of arithmetic (this system may be Peano's or any system of at least equivalent strength). A contains at least one undecidable proposition (whose negation is also undecidable).*

When Gödel first presented this theorem, very few in the audience were able to grasp its philosophical significance, with the exception of first-rate mathematicians such as John von Neumann. The very next day, von Neumann sent him a note praising his work and demonstrating an important corollary. This corollary was already known to Gödel. It is stated as follows:

Theorem 4 (Gödel's Incompleteness Theorems)**.** *The proposition asserting that arithmetic is non-contradictory is a proposition that is true, but undecidable.*

Far from being a mere logical curiosity, Gödel's theorems mark the failure of Hilbert's project to axiomatize arithmetic. What we observe, in effect, is a divorce between truth and provability. Not everything that is true is necessarily provable within the axiomatic framework. It is as if, after the long detour of modern science, we are led back before that peculiar kind of judgement identified by Plato — judgements that are true, but lack justification.

The incompleteness theorems seem a fruitful starting point for reconstructing Kant's definition of the synthetic *a priori* judgement. They argue, indeed, for a notion of truth that is not analytic, in the sense that it cannot be proven from axioms. Yet this truth still concerns pure objects of thought. It applies to the numbers of arithmetic, whose existence does not seem to depend on empirical testing in the world.

It is as though there existed two kinds of concepts: (1) abstract judgements rationally defined using mathematical symbols and axioms, and (2) intuitive judgements which are true without being rationally provable. The former concern rational concepts; the latter concern a kind of concept one might call intuitive. As an example, the intuitive concept of a whole number is what you think of when you enumerate objects, when you count them,

when you compare their quantities. The rational concept of a whole number is defined by a set of axioms such as the following:

1. The element called zero, denoted 0, is a natural number.
2. Every natural number n has a unique successor, denoted $s(n)$, which is also a natural number.
3. No natural number has 0 as its successor.
4. Two natural numbers with the same successor are equal.
5. If a set of natural numbers contains 0 and contains the successor of each of its elements, then that set is N.

Ideally, we would like the rational definition of the concept to match exactly our intuitive notion of number. In the case of Euclidean geometry, this is indeed what happens. All the properties one observes when drawing figures with pencil and compass can be proven using a wisely chosen set of axioms. This perfect correspondence between intuition and reason, between manipulable models and axioms, between intuition and formal logic, is what is called the completeness of geometry. And this completeness is bidirectional: every intuitive truth about figures is provable, and everything provable is intuitively true.

What holds in geometry, however, does not seem to hold in arithmetic. The greatest error in Platonic philosophy may have been to see, in the special case of circles and triangles, the absolute model of rationality. The Socratic quest — which searches, which tracks, which hunts down behind every idea a rational essence — resembles the quest for the Holy Grail. After two thousand years of hesitation, Gödel's theorems have brought the ecstatic paradise of pure ideas back down to the solid ground of disillusion. From now on, it is no longer permissible to say that intuitive and rational concepts are simply two sides — two mocking faces — of the same sestertius. On the contrary, one must admit that there are intuitive concepts that cannot be rationalized, and that there are rational ideas that correspond to no human intuition. It is by leaning on this idea of a fundamental complementarity between representation and definition that I will reconstruct the concept of the synthetic *a priori* judgement in Chapter 8.1, devoted to intuitive concepts.

5.2.5 Truth is a Sieve

Before rushing to conclusions, it is legitimate to wonder whether the interpretation I have offered of Gödel's theorems might not be overly charitable. One might think that true yet unprovable propositions are, in the end, mere

marginal cases within arithmetic. Gödel's sentence P, after all, is nothing more than a contentless sentence, speaking of itself in the void of logical space. It says nothing substantial about mathematics. Perhaps, indeed, this sentence could be eliminated? One might ask whether the incompleteness identified by Gödel could be circumvented simply by adding more axioms. I will summarize these objections below in the form of three major questions:

1. Is it possible to complete arithmetic by adding Gödel's undecidable proposition to the axioms of the theory?

2. Do the undecidable propositions identified by Gödel's theorems carry any arithmetical significance, or are they merely vacuous phrases?

3. Are there other undecidable propositions beyond those uncovered by Gödel?

Objection (1) is based on the idea that whenever an undecidable proposition is encountered, it suffices to stipulate within the axioms whether it should be held as true or false. Once included among the axioms, the proposition becomes trivially provable. We encountered this approach when analysing Euclid's *Elements*. We saw that the fifth postulate is in fact undecidable with respect to the other four; in other words, there exists no way to prove or disprove it. It follows that one can either treat the postulate as true or treat it as false. In the first case, one obtains Euclidean geometry. In the second, as we saw, one obtains non-Euclidean geometries.

Unfortunately, this artifice fails to resolve the problem. If Gödel's proposition P is added to the axioms A of arithmetic, we obtain a new axiom system A'. This system A' still satisfies the premises of Gödel's first theorem: it remains capable of formalizing arithmetic. It is therefore possible to reconstruct, within A', a new Gödel sentence P' which is true but undecidable. Incompleteness returns with force, like the tide surging against our modest sandcastles. But perhaps one will say that we must not yield to wind and dust? That we must repeat the reasoning unto death itself? Alas, such effort is in vain. For if we add once again this proposition P' to the axioms of the theory, we obtain a new axiom system A" from which yet another Gödel sentence P" will emerge, and so on. No sequence of choices will ever yield a complete system. Put differently, the axioms of arithmetic contain a hole that cannot be patched—unless one agrees to admit an infinite number of axioms.

Even if we are forced to admit the presence of Gödel's unprovable proposition, one might still try to downplay its relevance. Objection (2) claims that Gödel's proposition P, as compelling as it may be, holds no meaning for arithmetic. A proposition that states of itself "I am unprovable" appears more akin to a linguistic trick than to an arithmetical truth. This objection, which I sketch here in a few brief phrases, was already circulating within

the scientific community long before Gödel's theorems became known. The mathematician Henri Poincaré, to name but one, said of Russell's paradox: "Logic is no longer sterile—it gives birth to contradictions[15]."

To respond to Poincaré's statement, I would like to offer two remarks. The first is that Gödel's second theorem, far from being sterile, seems to say something crucial about arithmetic. It says, in substance, that it is impossible to prove that arithmetic does not contradict itself. This result, if we dare to look it squarely in the eye, is quite extraordinary. Since time immemorial, the idea that "$2 + 2 = 4$" seemed to flow so naturally from reason that the phrase became synonymous with logical evidence, an ideal of eternal truth, a paragon of deductive reasoning. And yet, I can already hear the murmurs of my detractors behind the curtain. Perhaps they will say that Gödel's second theorem is not an arithmetical truth *in itself*, but a truth *about* arithmetic—a meta-property.

To silence the chatter once and for all, and here comes my second remark, it must be emphasized that since the publication of Gödel's theorems, truly arithmetical undecidable propositions have been discovered. In 1970, Yuri Matiyasevich, building on prior work by Julia Robinson and Martin Davis, succeeded in showing that there exists no general algorithm capable of finding the integer roots of a polynomial. What do I mean by that? We know that certain equations like $x + 1 = 0$ admit integer (non-decimal) solutions. We also know that equations like $x^2 + 2y^2 = 0$ have no integer solutions. In the early twentieth century, the mathematician Hilbert raised the question of whether one could find a computational method capable of determining, without error, whether a polynomial equation with n unknowns and integer coefficients has any integer solutions. These polynomial equations with integer coefficients take the following form:

$$a_1x + a_2x^2 + ... = 0$$

where the coefficients $a_1, a_2, ...$ are integers (non-decimal). Thus, we exclude equations of the form:

$$\sqrt{x} + \frac{y}{z} = 0$$

This problem, known as Hilbert's Tenth Problem, received a negative answer following Matiyasevich's publications. It follows that, regardless of the resolution method you choose, it is always possible to isolate at least one equation for which your method will fail to provide a satisfactory answer in finite time. This result of undecidability leads to a particularly interesting reformulation of Gödel's theorem.

Theorem 5. *Every time we construct a set of axioms capable of formalizing arithmetic, it is possible to find a polynomial equation with integer*

[15] Henri Poincaré, *Les mathématiques et la logique*, p.306.

coefficients that has no integer solutions. This fact is true, but it cannot be proven within the given axiom system.

Let us note that Matiyasevich's results are not isolated. Since the 1940s, many undecidable problems have been discovered in the field of mathematics. One such example is the *mortal matrix problem*, which asks whether, by multiplying and adding together a set of $n \times n$ matrices with integer coefficients, one can obtain the zero matrix composed entirely of zeros. Another example is the *Post correspondence problem*. In this case, we consider two cells, each containing n distinct DNA strands. The question is whether each cell can juxtapose its DNA strands to produce the same final DNA sequence. It is thus a question of constructing the same object from different parts.

This growing number of undecidable problems brings us back to objection (3). This one, let us recall, questions whether unprovable propositions are numerous, or whether they remain a marginal phenomenon. One might think, after all, that undecidability is merely anecdotal—a behavioral oddity affecting the boundaries of mathematical knowledge, without threatening the general project of rationalizing intuitive truths through axiomatics.

As it turns out, this is not the case. In recent decades, substantial progress has been made in quantifying, in one way or another, the information that a given formal system can provide. This has led to a broader and deeper interpretation of the uncertainty theorems. A first attempt was undertaken by Chaitin[16]. He states a heuristic principle according to which the theorems provable by a given formal system cannot be significantly more "complex" than the theory itself.

Calude and Jürgensten have shown that for any formal system S, there exists a constant κ such that if the complexity of a logical formula exceeds κ, then the formula is unprovable[17]. The notion of "complexity" must here be understood in a rigorous sense. It is linked to the minimal number of symbols required to express the formula in a formal language.

An interesting question at this stage is: what proportion of propositions can a theory actually prove? To answer this, the authors begin by showing that there exists an infinite number of propositions ϕ such that $\delta(\phi) > \kappa$, or in other words, there exists an infinite number of propositions that are not theorems of the theory. Continuing this line of reasoning, they conclude that the longer a proposition is, the more likely it is to satisfy this condition. In the limit case where their size n tends to infinity, almost none are theorems. This leads them to a generalized form of Gödel's theorem, which I personally call *The Insular Theorem*. It states that among all the conceivable mathematical truths, only a small island—swept by the winds—is provable. Provable formulas are like the natural numbers: there are infinitely many of them, but the closer one gets to infinity, the scarcer they become. At a

[16] See Calude and Jürgensten [**?**], *Is complexity a source of Incompleteness?*, p. 1.
[17] *Ibidem*, p. 9.

certain point, they are no more than isolated particles in the freezing cold, the gaping void of vast distances.

Theorem 6 (Calude and Jürgensten, 2005)**.** *Let a formal system satisfy Gödel's assumptions. The probability that a true proposition ϕ of size n is derivable within the theory tends toward zero as n tends toward infinity.*

If one is allowed to summarize the current state of the logicist project in a few words, I would say that the contemporary situation aptly illustrates the tension between analytic propositions and intuition. Far from precisely capturing the intuition behind mathematical theories, the development of a rigorous axiomatics seems, with each passing day, to reveal more clearly the relative autonomy of logical thought from intuitive thought. It is as though the human mind were capable of conceiving totalities that logic itself is incapable of expressing.

The greatest logical paradox, that being said, may lie in the fact that despite the profound impotence of calculative procedures, and the ultimately very limited number of truths that are rationally accessible, we persist nonetheless in using calculative reason and the computer, as if these were the only legitimate paths to knowledge.

5.3 Theories of Computability

We now come, incidentally, to the second paradigm of nascent computer science: the axiomatization of rational reasoning through the construction of a model of automatic computation. This model is to function in such a way that, for any question asked—provided that the question is sufficiently precise to receive an unambiguous answer—a valid answer may be obtained. Reason, from this perspective, may be modeled like a black box that associates an input with an output in a deterministic fashion.

To define in precise terms the concept of computability, Hilbert referred to the fiction of a man deriving symbols across a stack of sheets. If we take this conception seriously, but abstract away tendons, nerves, and all the contingent biological framework that constitutes a human being, we may reduce the idea to its most elementary form: on the one hand, a paper tape divided into a number of discrete cells; on the other hand, a mechanical reading head able to move freely along the tape, stop on a cell, read the symbol written there (or detect the absence of a symbol), erase it, and write a new one in its place.

Any computation would then consist, for the machine so defined, in reading a question supplied to it on the tape using a certain number of symbols, and offering an answer on the same tape with similar symbols. The kind of admissible symbols is defined by an alphabet Σ. In its simplest form, this alphabet is binary and thus composed only of 0 and 1. In its

most sophisticated form, the alphabet may contain an infinite number of symbols. One need only think, for instance, of the set $\mathbb{N}$ of natural numbers (countably infinite), or the set $\mathbb{R}$ of real numbers (uncountably infinite). Whatever the nature of the alphabet, the symbol $\sqcup$ is used to represent the fact that a given cell on the tape is blank. The symbol $\sqcup$ is therefore a reserved symbol, or metasymbol, which cannot be included among the other symbols of the alphabet Σ.

If the tape is "sufficiently long," then any alphabet can be represented in binary form by the machine. Concretely, this means that the size of the alphabet does not affect the type of questions the machine is able to read, nor the type of answers it is capable of producing. By “sufficiently long,” I mean a tape whose length can be extended at will. Such a tape corresponds to the situation Hilbert envisioned when he said that the calculator has the right to use as many sheets of paper as he wishes. At any given moment, the number of sheets used can be described by a finite number, but this number itself knows no upper bound.

Another important property, aside from the length of the tape, is the memory of the machine. In Hilbert's hypothesis, the mathematician does not merely read symbols on a sheet of paper; he may also annotate those sheets during his reading—or even jot down his thoughts in a separate notebook. A machine that, like this human calculator, possesses the ability to retain elements in memory will be able to answer more complex questions than one that simply scans the document symbol by symbol without retaining anything.

Machines with finite memory are called *cellular automata*, and there are various kinds: those equipped with an external notebook, and those that are limited to reading a finite tape and answering "yes" or "no" at the end of the process. Machines with indefinite memory, on the other hand, which are further capable of writing responses more elaborate than a simple "yes" or "no", are known as *Turing machines*.

Although the use of computers has accustomed us to the idea that a machine's memory capacity is limited, I will show, at the end of this section, that the most adequate model for representing the notion of computability is the Turing machine. The argument splits into a minimalist and a maximalist side.

The minimalist side consists in exhibiting a function which, though uncomputable by a cellular automaton, is computable by a Turing machine (and would also be computable by a human with a sheet of paper as long as needed). We therefore need at least a model as powerful as the Turing machine to capture the concept of computability.

The maximalist side, for its part, consists in stating that there exists no finite model more powerful than the Turing machine. In other words, any conceivable modification aimed at improving the Turing machine—by equipping it, for instance, with additional memory tapes or more sophisti-

cated reading heads—may indeed increase its speed of execution, but not its actual computational power. The Turing machine is thus the smallest model that fully captures all the computational capacities of a computer.

It is important here to note the distinction between a real machine, like a computer, and the Turing machine. The computer is not a mathematical model but a physical object equipped, in addition to a certain structural organization, with wires and electrical signals (hardware). To model the computer, one would have to use a Turing machine with a finite-length tape. This would occupy an intermediate place between the cellular automaton—which also has a finite tape but cannot modify it—and the true Turing machine.

The Turing machine with an infinite tape, on the other hand, is not a model of the computer but of the abstract notion of computability. It is a model of computability in the same way that the mathematical triangle is a model of geometry. There are probably no perfectly triangular solids in the world, as Descartes once remarked, just as there are no computers with truly infinite memory. Nevertheless, the study of triangles teaches us something about the roughly triangular solids that we encounter in the world. In the same way, studying the properties of a Turing machine informs us about the powers and limitations of computational processes—and thus, indirectly, about certain properties of the human mind (assuming that the human mind consists, at least in part, of something computable).

5.3.1 Formalism of the Turing Machine

Alphabet

Before going any further, it is necessary to lay down a few formal definitions. A set of symbols Σ is called an alphabet, whether it is finite or not. The most common alphabets are the binary alphabet $\{0, 1\}$ and the Latin alphabet $\{a, b, c, ..., y, z\}$. There are, however, no restrictions on the symbols that Σ may contain. Often, the character $\sqcup$ is used to represent a blank space.

A *string of characters*, also called a *word*, is a sequence of symbols drawn from this alphabet. It can be represented as a sequence $\sigma_1, \sigma_2, \sigma_3, ...$, where σ_1, σ_2, and σ_3 belong to Σ. Thus, we may say that 100101 is a string of characters from $\{0, 1\}$, and that *logique* is a word from the Latin alphabet. Under this convention, the symbol 0 represents both a single element of the alphabet and the string composed of that single element. Moreover, we denote by "" the empty string.

The length of a string s is the number of symbols it contains. It is often denoted $|s|$. For example, $|machine| = 7$. When needed, we write $s(i)$ for the symbol at the i-th position of the string s, so that in the example above, $s(1) = m$ and $s(6) = n$.

Two strings x and y defined over the same alphabet can be concatenated

using the "$\cdot$" operator. The result of this operation is the word obtained by appending the symbols of y to the right of x, while preserving their order. In other words, $100101 = 10 \cdot 0101$ and $parapluie = para \cdot pluie$. When the context is sufficiently clear, the concatenation symbol may be omitted, and we simply write xy instead of $x \cdot y$. In either case, x is called the prefix and y the suffix.

The reverse of a string x is written x^R. It consists in writing each of its symbols in the opposite order. The reverse of *repas* is thus *saper*. In the special case where $x = x^R$, we speak of a palindrome, such as *ressasser*, *radar*, or 1001.

Language

Given an alphabet Σ, we denote by Σ^* the set of all words that can be formed in this alphabet by concatenation of its symbols (the operation $*$ is known as the Kleene star). In the case of the Latin alphabet, Σ^* contains all the possible or imaginary words that can be constructed. Most of them are unpronounceable, and their length would rival that of the mythological serpent Jörmungandr, but a small minority of these words belong to the French language. From what we have just said, a *language* is thus a subset of the strings, or words, that can be formed from a given alphabet.

Whereas the French language contains a finite number of words, which can be listed in a dictionary, this is not the case for most languages studied in computability theory. For instance, the language L of palindromes is the infinite set of words that can be read both from left to right and from right to left. One writes $L = \{x \in \Sigma^* \mid x = y \cdot y^R, \text{ with } y \in \Sigma^*\}$. In this expression, the property $x = y \cdot y^R$ is what distinguishes palindromes from other words in Σ^*.

Let us note in passing that the very act of constructing a language L from a property, without specifying the procedure that would allow one to enumerate the words of this language methodically, is what Poincaré called extensive infinity. Infinite, because such languages are often of infinite size; extensive, because the resulting language is, quite simply, the extension of the words it comprises.

It is easy, by relying on extensive infinity, to produce paradoxes. To do so, suppose first that all the words formed from the Latin alphabet Σ, to which we add the blank space $\sqcup$, are sorted in lexicographic order. This ranking would begin with one-letter words starting with a, then with b, then with c, and so on; followed by two-letter words starting with a, then with b, and so on. Now let us define a language L consisting of all words with the following property: "The word appears before the first word in lexicographic order that cannot be expressed in fewer than 132 characters."

We observe that the sentence in quotation marks contains fewer than 132 characters, spaces included. Consequently, either all the words in our

lexicographic ordering are expressible in fewer than 132 characters, in which case L contains all possible words and their number is finite — which contradicts the idea that the set of possible words is the countably infinite concatenation of the alphabet's letters; or, upon enumerating the words in lexicographic order, one eventually reaches a word that cannot be expressed in fewer than 132 characters — but that word is in fact expressible, precisely by the sentence in quotation marks. The language L is thus truly paradoxical[18].

What renders this language paradoxical is that the property P on which it relies refers to a certain ordering of the words. But once the ordering is constructed, it becomes possible to reference it in order to obtain a new word using fewer words than were necessary in the original reference. This leads to a new reference, and so on, ad infinitum.

To avoid the emergence of such paradoxes, it may be preferable to consider only those languages that can be constructed within the alphabet using solely the operations of concatenation, Kleene star $*$, and the union of several such languages. These languages are called regular, and by extension, their words are called regular expressions. An example of a regular language is $\mathcal{L} = \{w \in \{0,1\}^* : w$ has two or three occurrences of 1, the first two of which are not consecutive$\}$. This language is regular because it is possible to state the rule for constructing its words using only the usual rules of concatenation and the $*$ operation. These are the words obtained as follows: $0^* \,.\, 1 \,.\, 0 \,.\, 0^* \,.\, 1 \,.\, \{10\}^*$.

When a language is regular, it is easy to determine whether a given word belongs to it. Otherwise, one may ask whether there exists an algorithmic procedure capable of determining whether a sequence of symbols is or is not a word of a given language. The existence of such a procedure guarantees that the language in question is constructive. In Hilbert's terms, this would mean that the language can be obtained through a certain manipulation of symbols. The simplest manipulation we can imagine, at this stage, is that of the cellular automaton.

Cellular Automaton

A cellular automaton is a reading head capable of reading the content of a tape, one cell at a time, and responding at the end of the reading with either "yes" or "no". In the first case, it is said that the tape content (the input) is accepted; otherwise, it is rejected. Such a machine cannot move backward with its reading head, nor change the content of the tape. However, as we will see, it allows us to determine whether a word belongs to a regular

[18] This paradox is a reformulation of Berry's paradox, introduced under that name by Bertrand Russell in honour of the eponymous London librarian. The paradox considers: "the smallest integer not nameable in fewer than eighteen syllables," which appears to be so named in seventeen.

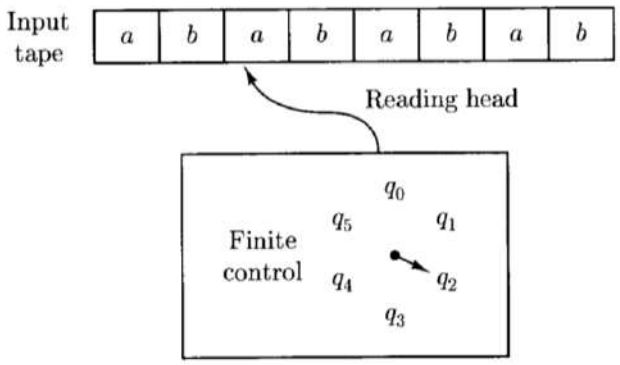


Figure 2-1

Figure 5.1: Diagram representing a cellular automaton. It has a tape, a reading head, and an internal algorithm determining the state the machine should be in. The tape is read only once, from left to right. (Image source: Lewis & Papadimitriou.)

language.

To formalize the cellular automaton, it must be equipped with a finite set of states $K = \{q_0, q_1, q_2...q_n\}$, as well as a set of instructions (an algorithm) specifying: "if the machine is in state q and reads a certain symbol σ, it moves to state q'". The instructions are thus complete injections from $(K \times \Sigma)$ to Q.

At the beginning of the reading, it is necessary to specify in which state the machine starts. This is called the initial state. Each time it reads a cell on the tape, the machine may change state depending on what is prescribed by the algorithm. When it reaches the end of the tape, it will be either in a terminal state, in which case it will respond with "yes", or in a non-terminal state, in which case it will respond with "no".

Définition 4. *A deterministic cellular automaton is a quintuplet* $M = (K, \Sigma, \delta, s, F)$ *such that*[19]*:*

1. K *is a finite set of states* $\{q_0, q_1...\}$*;*
2. Σ *is an alphabet;*
3. $s \in K$ *is the initial state;*
4. $F \subseteq K$ *is the set of final states;*

[19] Harry R. Lewis, Christos H. Papadimitriou, *Elements of the Theory of Computation*, p.57.

5. δ is the transition function, an injection from $K \times \Sigma \cup \{()\}$ to K.

Example 1. *Consider a cellular automaton that can determine whether a word w belongs to the language $\mathcal{L} = \{w \in \{a, b\}^* \mid w$ contains no occurrence of $b\}$. We need only consider two states q_0 and q_1, where q_0 is the initial state and the only terminal state. The transition function δ is given by the table below:*

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

In this example, the cellular automaton $M = (\{q_0, q_1\}, \Sigma, \delta, q_0, q_0)$ starts in state q_0 and reads the cells of the tape one by one. It transitions to state q_1 as soon as it encounters a b and never leaves this state. At the end of the reading, it responds "yes" if it is in state q_0, and "no" if it is in state q_1.

It is possible to represent the transition function δ of a cellular automaton using graph theory. Each node corresponds to a state, and the arrows going from one node to another represent the allowed state transitions. The label for these arrows is the condition required to perform the transition. Thus, the automaton from the above example can be represented using the following diagram:

The most important result concerning cellular automata is that they can recognize regular languages, and only regular ones. There is, therefore, a conceptual identity between a regular language and a cellular automaton.

Theorem 7. *A language is regular if and only if it is accepted by a cellular automaton.*

Proof. A sketch of the proof consists of showing that all operations intrinsic to regular languages can be accepted by a cellular automaton. Without loss of generality, consider a cellular automaton $M = (K, \Sigma, \delta, q_s, F)$ with $\Sigma = \{\sigma_1, \sigma_2, ...\}$.

- **Concatenation operation:** When the machine needs to check the concatenation of two letters $\sigma_1\sigma_2$, it enters state q_c. We construct δ so that $\delta(a_2) = q_{yes}$, but for any other symbol $\sigma \neq a_2$, we have $\delta(\sigma) = q_{no}$.

- **Kleene operation** $(^*)$**:** Let $\Sigma_1 \subseteq \Sigma$. When the machine needs to verify that the sequence of the word belongs to Σ_1^*, it enters state $q*$. We construct δ so that for any $\sigma \in \Sigma_1$, $\delta(\sigma) = q*$, while for any $\sigma' \notin \Sigma_1$, $\delta(\sigma') = q_f$.

- **Union:** Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two regular languages. Without loss of generality, we assume they are defined over the same alphabet Σ

(if not, we redefine these languages over the union of their respective alphabets). Let $M_1 = (\{q_1, q_2, ..., q_n\}, \Sigma, \delta_1, q_{s1}, F_1)$ and $M_2 = (\{k_1, k_2, ..., k_m\}, \Sigma, \delta_2, q_{s2}, F_2)$ be the machines accepting the words of these two languages, respectively. To obtain a machine accepting $\mathcal{L}_1 \cup \mathcal{L}_2$, we construct $M = (K, \Sigma, \delta, q_s, F)$. For each state q_i and each state k_j, we construct the conjugated state $q_{i,j}$ in K (with $i \leq n$ and $j \leq m$). The transition table for δ is obtained by conjugating the transition tables of δ_1 and δ_2. For any state $q_{i,j} \in K$ and any symbol $\sigma \in \Sigma$, the table thus returns the state $q_{k,l}$, where k is the index of state $\delta_1(\sigma, q_i)$ and l is the index of $\delta_2(\sigma, k_j)$. A state $q_{i,j}$ is final if and only if both q_i and k_j are final states.

This proves that for every regular language, a machine accepting that language can be constructed. To prove the converse, consider any cellular automaton $M = (K, \Sigma, \delta, q_s, F)$ and show that the language accepted by M is regular. Without loss of generality, assume $q_1 \in K$ is the initial state. Let $R(i, j, n)$ be the set of strings that can be read on the tape for the machine to transition from state q_i to state q_j without passing through an intermediate state with an index greater than k (thus allowing states $q_1, q_2, ..., q_k$). It is easy to see that the languages accepted by the machine are given by:

$$\bigcup_{j,n} R(1, j, n) : q_j \in F$$

We know that the union of two regular languages is regular. We will prove by induction on n that for all i, j, n, $R(1, j, n)$ is regular. To do so, we first note that $R(i, j, 0)$ is regular. This is quite obvious because it is either an empty language or a single character. Next, suppose that $R(i, j, k)$ is regular for i, j and for all $k \leq n$. The only strings allowing the machine to go from state q_i to state q_j without passing through more intermediate states with indices greater than $k + 1$ are:

- Either from state q_i to q_j without passing through an intermediate state with an index greater than $k - 1$;
- Or from state q_i to q_k, then zero or more times from q_k to q_k, and then from q_k to q_j, all without ever passing through a state greater than k.

Thus, we can write:

$$R(i, j, k + 1) = R(i, j, k - 1) \cup R(i, k, k).R(k, k, k - 1)^*.R(k, j, k - 1)$$

Since the concatenation and Kleene union of regular languages yield regular languages, we conclude that $R(i, j, k + 1)$ is regular for all indices i, j and k. □

As an illustration of this theorem, it is easy to see that there is no cellular automaton capable of accepting the language of palindromes $L = \{x \in \Sigma^* \mid x = y \cdot y^R, \text{ with } y \in \Sigma^*\}$. Indeed, the machine would need to memorize the letters it reads as it goes, to check if, once it reaches a certain point in its reading, the sequence of memorized letters repeats in reverse order. To achieve this, one might think of equipping the automaton with many states $q_a, q_b, q_c, ...$ corresponding to cases where the first letter read is an a, a b, a c, as well as states $q_{aa}, q_{ab}, q_{ac}, ..., q_{ba}, q_{bb}, q_{bc}...q_{ca}, q_{cb}, q_{cc}...$ allowing it to record what the second letter read was, and so on. But the problem with this approach is that the number of states must be fixed in advance. If there are n states and the word to be read contains too many letters, there will not be enough states to record all possible situations.

Thus, we understand that palindromes of indeterminate size do not form a language accepted by a cellular automaton, and therefore, are not a regular language. This is consistent because the operation y^R is not a simple concatenation operation nor a simple Kleene * operation. It seems natural, then, to wonder what extension should be added to the cellular automaton for it to accept palindromes.

Stack Automaton

A simple extension of the cellular automaton, capable of recognizing palindromes, would consist of adding an auxiliary tape to the machine. This tape would be used, at any time during the operation, to store information and read it, without editing it. The result of this addition is soberly named the Stack Automaton, where the term "stack" is a translation of the English word and refers to the ability to stack information. Nothing to do with lithium batteries, therefore!

Définition 5. *A stack automaton is a six-tuple $M = (K, \Sigma, \Gamma, \triangle, s, F)$ such that*

1. *K is a finite set of states $\{q_0, q_1...\}$;*
2. *Σ is the alphabet of the tape;*
3. *Γ is the alphabet of the stack;*
4. *$s \in K$ is the initial state;*
5. *$F \subseteq K$ is the set of final states;*

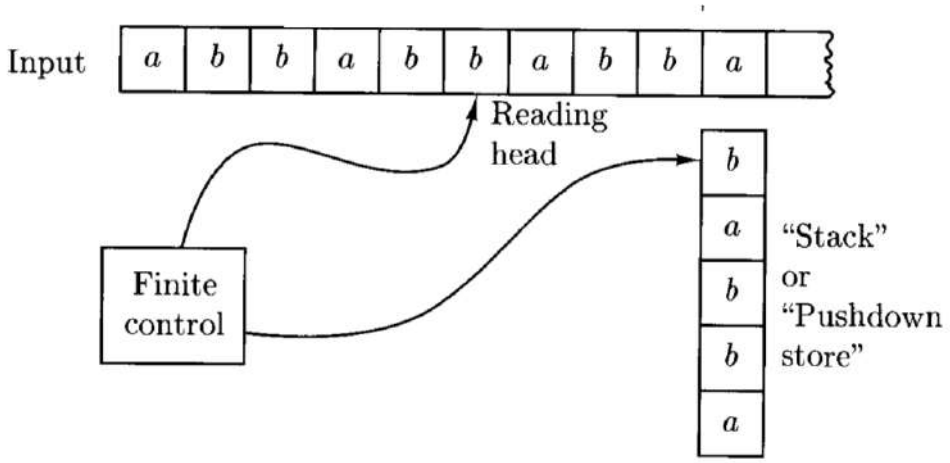


Figure 5.2: Diagram illustrating a stack automaton. This enhanced cellular automaton is equipped with an external tape where it can add or erase symbols. The addition of a symbol is always done on top of the existing symbols (the "pop" operation), and the removal consists of deleting the last written symbols (the "push" operation). (Image source: Lewis & Papadimitriou.)

6. *δ is the transition function, an injection from $K \times \Sigma \cup \{""\} \times \Gamma$ into $K \times \Gamma^*$.*

The function $\triangle : (q, \sigma, \gamma) \to (q', \gamma')$ means that, if the machine is in state q and reads symbol σ, it must transition to state q'. Additionally, it replaces the string γ at the top of the stack with the string γ'. In particular, the transition $(q, \sigma, "") \to (q', a)$ means that the symbol a must be added to the top of the stack; this operation is called "popping" the symbol a. Conversely, the transition $(q, \sigma, a) \to (q', "")$ means that the symbol a must be removed from the top of the stack; this operation is called "pushing" the symbol a. This terminology justifies the alternative name of "pushdown automata" for stack automata.

The Turing Machine

In addition to the reading head and the finite tape we discussed earlier, a Turing machine is also equipped with an *algorithm* that defines, at each unit of time, the operation to be executed. The allowed operations are: moving the reading head one cell to the right ($\to$) or one cell to the left ($\leftarrow$); reading the symbol $\sigma \in \Sigma$ on the current cell (denoted $.\sqcup$ if the cell is empty), or writing a new symbol $\sigma' \in \Sigma$.

The algorithm is essentially the machine's program. It consists of a number of $if..., then...$ instructions. For example: "If the tape is on the first cell, then move to the second cell." Obviously, we do not want the machine

to always move from the first to the second cell. Therefore, we will specify: "If the reading head is in state q_0 and is on the first cell of the tape, then move to the second cell; but if it is in state q_1, then read the symbol on the cell, and so on." The number of states $q_0, q_1, q_2...$ allowed by the machine is unlimited but bounded. It is thus impossible to have an infinite number of states, just as it is forbidden to have an algorithm with an infinite number of instructions.

In addition to the initial state, the algorithm also prescribes a final state q_f. When the machine enters this state, it stops immediately. The only instruction associated with q_f is therefore: "If the machine enters a final state, regardless of the position of its reading head, it stops." An important question is whether, for a given algorithm, the questions written on the tape have a correct answer in finite time. For this to be true, the machine must both stop and the string of symbols written on the tape when it stops must be the answer to the question posed.

It is easy to see that some machines stop for certain inputs and do not stop for others. For example, if you construct an algorithm for performing Euclidean division, the machine will stop for inputs like "divide 8 by 2", but not for "divide 10 by 3". In the latter case, the result on the tape will be 3.33333..., with an infinite sequence of consecutive numbers. However, it is important to note that if the question is "Is 10 divisible by 3?", it is possible to write a program that answers this kind of question in finite time. This program simply iterates through all the multiples of 3 until it reaches 10 (in which case 10 is divisible by 3) or a number strictly greater than 10 (in which case 10 is not divisible by 3). This can be written as a program:

```
Divisible(s1, s2)
    set i = s1
    while i < s2 :
        if i = s2, answer true ;
        else, set i = i + s1 ;
    else answer false;
```

Binary questions, whose answers are either true or false, seem easier to solve than open-ended questions. In technical terms, the problem of determining whether a certain property is true or false is called a *decision problem*. One example of a decision problem we have already encountered is determining whether an arithmetic proposition can be proven from a given set of axioms. We saw that the answer was negative in the formalism of Gödel's theorems. It is interesting to ask whether, in the formalism of Turing machines, all decision problems have a solution.

A string of symbols is said to be accepted by the machine if, when this string is written on the tape, the machine eventually reaches a final state. Conversely, the string is not accepted if the machine never halts.

The set of strings $x*$ accepted by the cellular automaton constitutes the computable language of that automaton. These considerations lead us to a formal definition:

Définition 6. *A Turing machine is a quintuplet* $M = (K, \Sigma, \delta, s, H)$ *such that*

1. K *is a finite set of states* $\{q_0, q_1...\}$;
2. Σ *is the alphabet of the tape, containing the blank symbol* $\sqcup$ *as well as the symbol representing the beginning of the tape* $\triangleright$ *but not the symbols* $\rightarrow$ *and* $\leftarrow$;
3. $s \in K$ *is the initial state;*
4. $H \subseteq K$ *is the set of final states;*
5. δ *is the transition function, an injection from* $(K - H) \times \Sigma$ *into* $K \times (\Sigma \cup \{\leftarrow, \rightarrow\})$.
   (a) *For all* $q \in (K - H)$, *if* $\delta(q, \triangleright) = (p, b)$, *then* $b = \rightarrow$.
   (b) *For all* $q \in (K - H)$ *and* $a \in \Sigma$, *if* $\delta(q, a) = (p, b)$, *then* $b \neq \triangleright$.

5.3.2 The Halting Problem

A major question is whether all problems are decidable. A problem is said to be "decidable" if a Turing machine can answer it as either true or false using an unequivocal computational method. For instance, the question of whether a number is prime is decidable. There does exist an algorithm to determine it: this algorithm iteratively checks whether a given number s is divisible by any number strictly smaller than s except the number 1[20].

One nonetheless expects there to be undecidable problems. And for good reason: the Church-Turing thesis states that Turing machines and Gödel's axiom systems are two equivalent ways of expressing the concept of computability. They are thus two competing definitions of the same intuitive concept. This thesis, of course, is not a matter of proof. It is merely a consensus within the mathematical community.

It so happens that Alan Turing succeeded in determining the existence of an undecidable problem. This problem is known as the "halting problem." In intuitive terms, it consists in constructing a Turing machine, called *Halt*, which analyzes the behavior of all other possible and imaginable Turing machines in order to determine under what conditions these machines halt after a finite number of operations and under what conditions they loop infinitely. Our machine should therefore be capable of recognizing that a

[20] Donald Knuth provides a more elaborate algorithm in his book titled *The Art of Computer Programming: Fundamental Algorithms*, Addison-Wesley, 1997.

program performing Euclidean division will halt when attempting to divide 10 by 2, but not when dividing it by 3.

Creating $Halt$ is not at all straightforward. $Halt$ must in fact understand what the machines it analyzes are doing in order to determine whether they have entered into an endless process. This understanding cannot be obtained simply by looking at the symbols these machines write on their reading tapes. If we observe a program, for example, that writes the following numbers 1421421421..., we might think it will never stop repeating itself until the end of time. Yet it may well be that, after 100 steps, the program halts or completely shifts course. To be sure, a human being would consult the machine's code. If the code takes the following form, one would know that the program halts after a finite amount of time:

```
procedure T(s):
    write 142;
    repeat one hundred times;
```

If, on the other hand, the code is written as follows, one would know that the program continues indefinitely:

```
procedure T(s):
    write s;
    if s is even, divide s by two;
    if s is odd, multiply s by three and add one;
    if s is equal to 100, halt;
    otherwise repeat;
```

In this latter code, the machine halts only if the number s at some point becomes equal to 100. This can happen when it starts with an input $s = 100$ or $s = 200$. However, there are cases where the program will never reach 100. It is easy for a human to verify that as soon as s reaches the value 1, the program merely repeats 142142... forever. In mathematics, this sequence of numbers even has a name: it is called the Collatz sequence. If you wish to gain money and renown as a mathematician, all you need to do is prove that this program will always reach the value 1 or 100 after a finite number of steps, regardless of the initial value of s. It is indeed an unsolved problem in number theory.

From a formal point of view, the halting problem, identified by Alan Turing, therefore consists in constructing a Turing machine $Halt$ capable of analysing the semantics of a program. It must grasp what the program is attempting to do and use its mathematical knowledge to predict the outcome of the process. $Halt(T, s)$ takes two arguments: the name of the machine T to be analysed and the input s to that machine. T and s are thus both strings. For convenience, it is assumed that two Turing machines bearing

the same name T always have the same code. To prove that $Halt$ does not exist, one possible approach is to assume that such a machine does exist and then show that this leads to a contradiction. This proof strategy is known as proof by contradiction — *reductio ad absurdum* in Latin.

Rather than delving into the technical details of the proof of $Halt$'s impossibility, I have imagined a humorous parable. I believe it may help you develop an initial intuition about the problem. Just listen: a long time ago, in a forgotten valley, there lived a master magician by the name of Tigran Ulan (an anagram of Alan Turing). Once a day, he would gather his disciples on the hill and, after pronouncing a word aloud, ask them to empty their minds. The master was well aware of the difficulty of this exercise, especially when the word spoken aloud was "Bella" — a woman from the neighboring village who stirred many a fantasy. There was no time limit allotted to the exercise, so after an hour had passed, it became difficult to say whether those who had not yet succeeded would ever manage to do so.

To address this problem, the Master fortunately possessed a divination spell, named $Paradoxe$, which allowed him to predict which of his students would succeed in the exercise after a finite amount of time and which ones, even if given an eternity, never would. The working of $Paradoxe$ followed the principle of communicating vessels: when the magician cast this spell on someone whose mind would never come to rest, he himself would instantly enter a state of absolute peace; whereas, when he cast the spell on someone whose mind would eventually come to rest, the Master's own mind would begin to accelerate, endlessly repeating the name "Bella." This effect was somewhat unfortunate, but the Master had grown used to it. He would often say: "Appearances are deceptive — the last shall be first!"

One day, a disciple of the magician, frustrated at his own failure, approached the Master and said: "Master, why do you not do the exercise with us? We want to know whether you, too, can resist Bella's charms!" The Master had trouble hiding his embarrassment. After spending so much time reading the minds of his students, had he not come to fantasize about the innocent maiden himself? Fortunately, he had an escape. "If I cast $Paradoxe$ upon myself," he explained, "should I succeed or fail at the exercise?"

The disciple thought for a moment. He quickly realized that if the spell predicted something about the Master's mind, it would bring about the opposite. For example, if the Master's mind were to come to rest, $Paradoxe$ would agitate it endlessly. Was this not in contradiction with the spell's prediction? On the other hand, if the Master's mind were not to find peace, $Paradoxe$ would halt it instantly. The spell had truly earned its name!

In this parable, the minds of individuals correspond to Turing machines. The word "Bella" spoken aloud is the input s, and the cessation of mental activity represents the situation in which a Turing machine halts. The spell

Paradoxe can thus be represented by a program whose function is to predict whether a Turing machine will halt on input s. If it will, *Paradoxe* loops infinitely. Otherwise, *Paradoxe* halts. This prediction, naturally, requires the use of the function *Halt*. The code can therefore be written as follows:

```
procedure Paradoxe(x):
    if Halt(x, x) returns yes,
        infinite loop;
    otherwise,
        halt;
```

It is important to note the presence of the expression $Halt(x, x)$. This line states that the letter x can simultaneously refer to a program and to the input s given to that program. This condition is met by the fact that each program can be represented by a name and that two programs with the same code always bear the same name. Aside from this technical detail, *Paradoxe* behaves just like in the parable I gave.

It is easy to see that *Paradoxe* behaves paradoxically when we ask it to run on itself. Suppose, indeed, that $Paradoxe(Paradoxe)$ halts. Then $Halt(Paradoxe, Paradoxe)$ will return yes, $Paradoxe(Paradoxe)$ halts. But in that case, *Paradoxe* must enter an infinite loop, which contradicts the assumption that it halts. Conversely, suppose that *Paradoxe* does not halt. Then $Halt(Paradoxe, Paradoxe)$ will return no, $Paradoxe(Paradoxe)$ does not halt. But in this scenario, the program *Paradoxe* halts, contradicting the assumption that it was not supposed to. The only valid conclusion is that there exists no machine *Halt* capable of telling when a machine halts.

5.3.3 Rice's Theorem

The second observation is a direct corollary of the halting problem. If there exists no machine capable of telling, for any given program, whether it halts on some given input, it follows quite readily that no machine can determine whether a program possesses any given non-trivial semantic property. By a *non-trivial property*, I mean a property that is neither true nor false for all possible and imaginable Turing machines. By a *semantic property*, I mean a property that pertains not to the lines of code as strings of characters, but to what the code is capable of doing. The question of whether a program contains more than ten lines of code, or whether it includes the word "then," is not semantic. Conversely, the following questions all constitute non-trivial semantic properties:

- Does the machine, at some point in its process, write the symbol 0 on the first square of its reading tape?

- Do the outputs of the machine $T(s)$ correspond to the multiplication of s by 2?
- Do the outputs of the machine $T(s)$ correspond to some given computable function $f(s)$?

The reason these properties cannot be decided by any algorithm is rooted in a phenomenon I shall now describe, known as semi-decidability. To verify whether a machine writes the symbol 0 on the first square of its tape at some point during its execution, one could simply observe that square and see whether, eventually, a zero appears. This method is perfectly effective for affirming that a program *does* indeed write a 0 in that location. But if, after, say, an hour of observation, nothing has happened, it becomes unclear whether the machine will *never* write a 0, or whether one need simply wait a little longer.

In this regard, we find ourselves in much the same position as the earliest humans: seeing the sun set each evening, they would ask themselves whether it might one day fail to rise again the next morning. In the absence of a law of nature governing the movement of celestial bodies, this question is undecidable. It is equally uncertain, as David Hume already pointed out, whether the laws of nature themselves may not encounter some unforeseen exception. One shall never know whether it might suffice to wait just one more morning to witness, at last, the absence of the sun. In technical terms, we say that a property is semi-decidable if it is possible to verify when the property *does* hold, but not to determine when it *does not*.

One might think it possible to circumvent this difficulty by constructing a sufficiently intelligent algorithm — one that would not merely inspect what is written on a given square, but would examine the program's very structure. This is precisely where the halting problem reemerges. If the halting problem were decidable, one could employ $Halt$ to determine whether a Turing machine $T(s)$ executes a certain function $f(s)$, such as the multiplication of s by two. It would suffice to first check whether the machine halts. If it does not, then it does not compute $f(s)$. If it does halt, then one need only verify whether the symbols written on the reading tape when the program halts do, in fact, correspond to $f(s)$. It follows that it is impossible to construct a general algorithm capable of analysing and determining the semantics of an arbitrary program. To bypass this impossibility, one must:

- Resort to case-by-case reasoning;
- Proceed by approximation;
- Restrict semantic analysis to a subset of key programs, making, for instance, specific assumptions about the structure of these programs.

The moral of Rice's theorem is that the analysis of a program's semantics is intrinsically difficult. Turing machines do not seem particularly adept at grasping the meaning of their own lines of code — or those of another machine. They are blind executors, manipulators of symbols, assembling them, dividing them, transforming them, without necessarily possessing a general overview, a power of self-inspection over the operations they are performing. They are mechanical arms, wires, sequential processes — devoid of *understanding* in the most demanding sense of the word. Perhaps never before have we been more entitled than today to say that machines operate in an *automatic* fashion, with all the impotence and faint disdain that this word quietly conveys.

5.3.4 Oracles and Cybernetics

The limitations imposed by Rice's theorem apply to all Turing machines, but not necessarily to every conceivable kind of machine. One might be tempted to ask whether it is possible to construct an extension of the Turing machine capable of answering undecidable questions—those pertaining, for instance, to the semantics of a program.

This kind of extension is not easy to come by. Alan Turing showed that most enhancements made to a Turing machine do not increase its computational power. For example, if one constructs a machine with multiple reading heads or several infinite tapes, this machine can do nothing more than a standard Turing machine. What I mean is that any function computed by an extended Turing machine T' can also be computed by an ordinary Turing machine T, although it may require more execution time. It is therefore possible to reduce the time needed to solve decidable problems, but not to resolve undecidable ones.

One way to bypass these limitations is to imagine an Oracle capable of "magically" solving any binary question in a single computational step. Much like the ancient Pythia or the Roman Sibyl, this Oracle can answer either "yes" or "no," with no risk of error. You may pose to it decidable questions such as, "Is the sum of the angles in a triangle equal to 180 degrees in Euclidean geometry?"—but you can also ask, "Does a given machine T halt on a given input s?" The trick behind this notion of an Oracle—aptly named, indeed—is that one need not describe how it functions. The Oracle never explains its answers. It is veiled in mist, wrapped in an aura of mystery, worthy of that phrase from Plutarch: "The Pythia speaks: many deny, some are ignorant, some deduce[21]."

To be precise, not all Oracles are necessarily capable of answering all questions. In certain contexts, an Oracle's responses may be restricted to particular classes of queries. For instance, one can design an Oracle that tells

[21] Plutarch, *Delphic Dialogues.*

whether a given input s for a machine T belongs to the set of inputs that halt in finite time. Such an Oracle solves the Halting Problem, though it may not solve other problems. One could also construct an Oracle that determines whether a given machine T actually implements Euclidean division—that is, whether its outputs match the expected result of such a division.

What is truly groundbreaking with the introduction of the Oracle is that we cease to concern ourselves with the internal mechanics of computation and focus solely on inputs and outputs. The Oracle can be likened to a mysterious "black box" (a Pandora's box?) that receives a question as input and returns the expected answer. This notion of an opaque "black box" stands in contrast to that of a fully transparent "white box." This terminology was first introduced by the mathematician Norbert Wiener, in a paper aimed at modeling communication among machines and animals (including human beings[22]):

> The terms "black box" and "white box" are convenient and figurative expressions of not very well determined usage. I shall understand by a black box a piece of apparatus, such as four-terminal networks with two input and two output terminals, which performs a definite operation on the present and past of the input potential, but for which we do not necessarily have any information of the structure by which this operation is performed. On the other hand, a white box will be a similar network in which we have built in the relation between input and output potentials in accordance with a definite structural plan for securing a previously determined input-output relation.

The answer provided by one black box may be fed into a white box, or into another black box. By linking boxes together, one creates a coherent information system where each stage of computation leads to the next. A humorous example of this approach can be found in a diagram by Marguerite Souchon, made popular on Twitter, which helps determine which Dostoevsky novel you should begin with in your exploration of Russian literature[23]. In this diagram, which I have attempted to reproduce in Figure 5.3, each box answers "yes" or "no" to a specific question. Depending on the answer, you are directed to a new question represented by another box. The process continues, thread by thread, until it culminates in a recommendation for a Dostoevsky novel.

Note that the process may have a cyclical behavior: it is permissible to reuse the final answer as a new input. This makes it possible, for example, to

[22] Norbert Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine*, Preface, page xi. (Footnote 1), 1948.

[23] Marguerite Souchon is a professor of Russian and the author of a book titled *Le Dieu de Dostoïevski*, published in 2021. On October 6, 2021, she posted a now-famous diagram on Twitter.

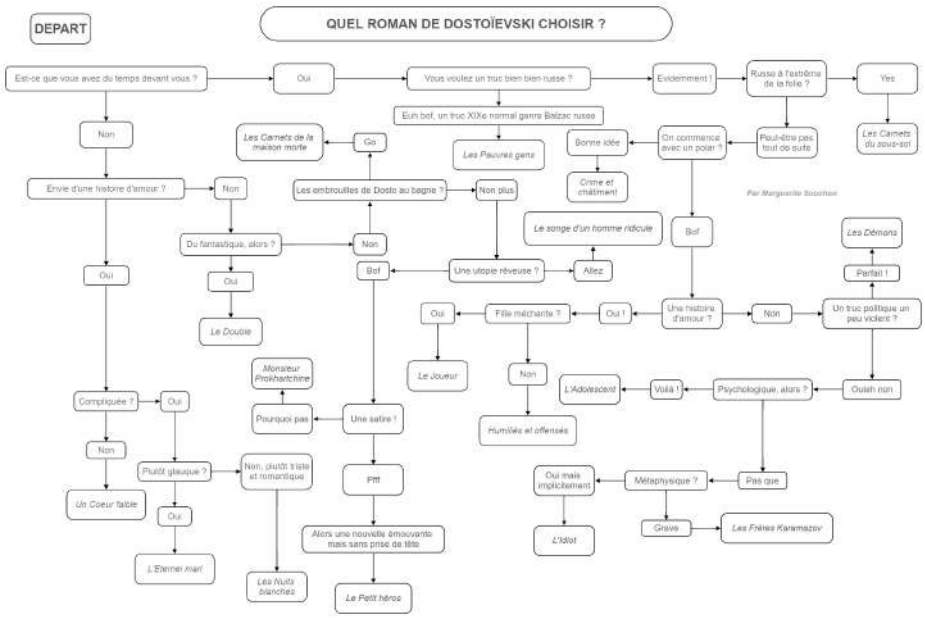


Figure 5.3: Diagram composed of connected black boxes allowing you to determine which Dostoevsky novel to begin with in your journey through Russian literature.

determine which book to read next, following the one you've just completed. This circularity is useful for suggesting new readings; or else, to revise your initial choice and guide you more wisely. This desire to use your answers to pose new questions—a procedure reminiscent of YouTube's recommendation algorithms—foreshadows a field of study known as *cybernetics*. Within this field, we no longer concern ourselves with machines composed of coils, functions, and computational procedures that yield a single result; rather, we study an ensemble of boxes interconnected by *feedback loops*. The answer from the final box is fed back into the initial box, correcting the system step by step until it converges on a refined answer.

A striking illustration of this feedback phenomenon is that of a hovering drone. If the system is well designed, the drone should remain stationary over time. Yet the forces acting upon it are so numerous that they constantly threaten to destabilize it. When the craft tilts to the left, the right turbine must be activated to restore balance; when it leans to the right, the opposite turbine is engaged, and so forth. This system can be represented as a set of black boxes that receive the drone's initial parameters (tilt, speed, position) as input and output the predicted future trajectory (tilt, speed, position). This prediction is then compared to the desired state (stationary flight). The discrepancy between the ideal and the predicted state is called the *error*. It is re-injected into the initial system in order to correct the trajectory at each

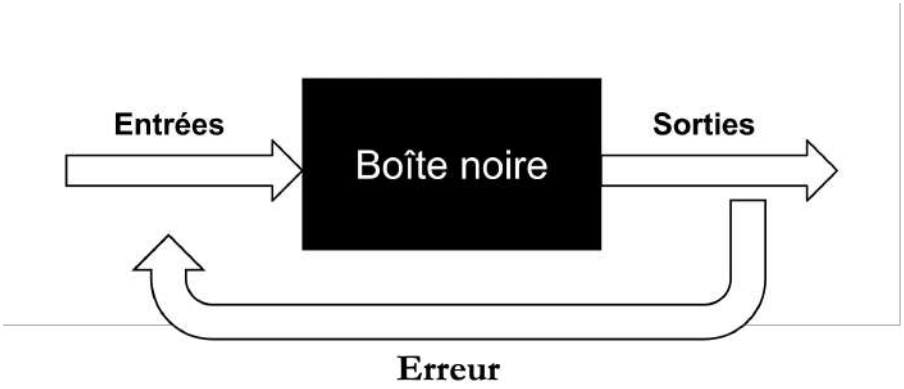


Figure 5.4: Diagram schematically illustrating the paradigm of computability. Instead of defining reason as a deductive method proceeding from axioms to conclusions, it may be represented as a black box that, for each input question, provides an output answer.

new computational step.

Note that in the example of the hovering drone, the black boxes do not respond to binary questions with "yes" or "no" answers. They respond to numerical queries such as: at what speed should the turbine be activated? at what angle should the drone tilt? and so on. Wiener's work can thus be seen as an extension of the Oracle concept. It is often considered the origin of cybernetic systems, defined as complex information systems.

Within the cybernetic paradigm, it matters little whether the black box provides its answers by means of a computable function or through mere statistical associations. What matters is that the answer be correct—or as close to correct as possible. When confronted with questions that are either too complex to compute or intrinsically uncomputable, it becomes tempting to resort to statistical models that yield plausible answers simply by mimicking known cases.

The most powerful statistical models currently available are neural networks. At first glance, these behave exactly like black boxes. They provide answers without explanation, striving to imitate known cases as faithfully as possible. We do not clearly understand the function linking input to output, and we are not even sure such a function exists in a manipulable form. What I mean by this is that there may be no general, simple, elegant, and humanly intelligible expression relating input to output. It is as though neural networks were the Oracles of modern computing.

In the next part of this thesis, I would like to suggest that the theories of provability and computability define two fundamentally different ways of thinking about programming. In the first, one works with proof systems based on axioms and logical laws, proceeding step by step; such systems can

be represented by white boxes. In the second, one deals with interconnected black boxes based on statistical data, whose behavior cannot be predicted *a priori*. These two perspectives—complementary in nature—represent the two faces of artificial intelligence. They form what I shall call *top-down* and *bottom-up* programming.

What to take away from this chapter

The abandonment of Kant's theory of judgement gave rise to Hilbert's formalism. In this line of thought, judgement is no longer a matter of intuition, but of pure logic.

More precisely, reason is defined as a certain capacity to manipulate symbols according to fixed rules, regardless of the meanings attached to those symbols. A good example of this is the solving of a mathematical equation, which requires manipulating algebraic letters while temporarily forgetting what those letters stand for.

- Hilbert's formalism is part of the broader idea that thought itself might be automated by a machine. Early computers, after all, are nothing more than devices for symbol manipulation.

  Yet this automation encounters certain obstacles: on the one hand, logic harbors paradoxes; on the other, there exist undecidable problems. As a consequence, some logical and mathematical truths cannot be discovered by means of unambiguous, top-down rules, but only through bottom-up methods.

  In the next chapter, I will show how this impasse gave rise to two fundamentally different paradigms of programming.

Chapter 6

Top-Down Approach: Expert Systems

Hilbert's dream was to find a method capable of determining, through a set of purely mechanical rules, what is true and what is false. If such a method existed, one could *de facto* do without human judgement altogether. Not only would there be nothing a machine could not do, in terms of scientific inquiry, but this machine would in fact be more reliable, more precise, and faster than we are. Yet, as we have seen, neither Frege's axiom-based approach nor Alan Turing's model based on the machine that bears his name allows us to totalize or fully automate knowledge. In other words, there will always be truths that are unprovable, whether by means of mathematical demonstration or computational calculation grounded in the Turing machine.

The failure of Hilbert's dream of automation is, for our purposes, a fortunate one. It suggests that the forgetting of the *Understanding* which characterizes the birth of the computer—and which, as we have seen, took place with the irruption of non-Euclidean geometries—faces certain theoretical limits. It does not seem possible to explain human knowledge solely on the basis of provability or computability. Everything points to the existence, between these two paradigms, of a third fundamental ingredient, one that was neglected by early computer science. Is it intuition? Is it a faculty of judgement?

The paradigm of provability relies solely on *syntax*. It seeks to formalize mathematical disciplines—starting with arithmetic—by reducing them to a set of axioms. A proposition such as "5 + 7 = 12", within this system, will be considered true if and only if there exists an inferential path leading from the axioms to the proposition. This path is not unique, but it is well-defined: one can trace its steps. What remains unknown, however, is what the machine capable of following these steps would look like. There is, to be sure, an abstract manual, but no concrete model.

Furthermore—and this is yet another difficulty—the incompleteness theorems of Gödel, as I have pointed out, suggest that the vast majority of propositions are unprovable. Axiomatic knowledge, obtained by purely analytical means, thus appears to be the exception rather than the rule. This island of reason, of clarity, of light—this paradise of Ideas praised by Hilbert, and before him by Plato—constitutes but a minuscule refuge within the cold, unfathomable, and enigmatic regime of human knowledge. How, on such fragile foundations, can we hope to embrace the vast expanse of understanding?

The paradigm of computability fares no better under critical scrutiny. It seems, at first glance, to escape Gödel's theorems. It adopts the standpoint of semantics. It is not so much concerned with understanding the logical rules of deductive reasoning as with describing, in principle, the workings of an automatic machine. This machine is equipped with a reading head, an infinite tape, symbols, and a mechanism as finely tuned as an authentic Charles Babbage engine[1]. It becomes possible to build it from scratch, to submit problems to it, to use it for encoding or decoding messages[2]. Yet what is gained on the practical side is lost in traceability. We would like to be able to perfectly predict everything a machine will do if a given input x is written on its tape, but the halting problem and Rice's theorem obstruct this dream of transparency. One must accept that the machine may exhibit emergent behavior, that certain responses may appear which were not previously foreseen.

There exists an epistemological tension between syntax and semantics, reminiscent of the one I described in the introduction to this thesis, between analysis and synthesis. Syntax seeks to understand the laws of reasoning. We look for axioms, we question the powers and limitations of logic. Everything is presented transparently. There is not a single result that cannot ultimately be justified by formal proofs. In this paradigm, human reason resembles a great juridical dictionary, equipped with laws, clauses, and a certain jurisprudence. There is no mystery, no surprise, no approximation, no ambiguity.

Semantics, by contrast, is less concerned with understanding the laws of reasoning than with simulating a machine capable of reasoning. This machine takes data and assembles them according to organic rules, to produce something new. Admittedly, its functioning is neither contingent nor

[1] Charles Babbage's machine, also known as the analytical engine, was a mechanical device intended to execute any computational program. It was never built, but the mathematician Babbage provided a detailed description. It is the first conceptual prototype of the computer.

[2] This is precisely what Alan Turing's team would later do to decipher Enigma, the machine used by the Germans during World War II. In doing so, they relied on the later work of Marian Rejewski and on electromagnetic machines nicknamed "bomba kryptologiczna".

mysterious. But to know what a Turing machine will do, one must run the program. If the result were known in advance, what need would we have for the program? The machine never does anything other than deterministic actions: it follows steps which, one by one, are predictable. It nonetheless resembles a laboratory experiment more than a mathematical proof. One provides it with an input and observes the result. One modifies the input and notes how it affects the response. It can be represented as a black box, which associates to every input a determined output.

When these two paradigms are employed in crafting artificial intelligences, they yield strikingly different outcomes. The syntactic and axiomatic approach envisions cognition through a logical-deductive lens. Computer programs are thus crafted as a series of instructions, to be read line by line. By meticulously following this manual, the machine is expected to deliver the correct response. Although the computer boasts unmatched efficiency and precision, the earliest forms of algorithms unearthed by archaeology date back to ancient Babylon. These are clay tablets detailing a recipe for a leek stew[3]:

```
Meat is not used.
You prepare water.
You add fat.
You add kurrat, cilantro, salt as desired, leek, garlic.
You pund up dried sourdough,
You sift it and you scatter it over the pot before removing it.
```

A more sophisticated and modern algorithm might involve determining whether a given integer n is a prime number. To accomplish this, one attempts to divide it by all integers between 2 and $n-1$. If successful, the procedure halts: it is not a prime number. If, after performing these divisions, no result is reached, we conclude that n is a prime number. Formalising this procedure yields the following:

[3]These instructions are recorded on a clay tablet dating back to 1700 BCE. Despite damage to the inscriptions, two recipes remain visible. The one I've written below (translated from English) and a recipe for roast lamb. See *Food in Ancient Mesopotamia: Cooking the Yale Babylonian Culinary Recipes* published in *Ancient Mesopotamia Speaks: Highlights of the Yale Babylonian Collection.*

```
Function prime(n):
        If n = 0 or n = 1 then return False
        For each i between 2 and n-1:
                If i divides n, return False
        Else return True
```

In the examples above, the rules are described in a *top-down* manner. The task to be accomplished—or the skill one seeks to model—is defined in advance. We begin with a specific problem, such as verifying certain properties of numbers, identifying a geometric object within a virtual environment, or modelling a hydrodynamic phenomenon, and from this starting point, designing involves defining a solution algorithm. It all unfolds like a cooking recipe: methods are devised to achieve a particular goal. The success of these methods depends on a number of assumptions; and as long as one remains within this given context, the results are assured. This method, also the first to emerge convincingly, is also known as goal-oriented programming, or *goal oriented*.

The semantic and synthetic approach, running counter to these recipe-book methods, seeks to build meaningful correlations between the inputs fed to the machine and its outputs. Rather than explicitly instructing it on the steps required to produce the dish, a set of parameters is left at its discretion: the temperature of the water, the quantity of each ingredient, the vigor with which the mixture is stirred, etc. The whole art of programming then consists in teaching the machine how best to select these parameters. The simplest strategy rests on the time-honoured Law of the carrot and the stick. The machine is asked to generate thousands of meals with random parameters and to retain only those that come closest to the desired recipe. By repeating this process multiple times—introducing variations between each attempt and keeping the most conclusive ones—we may hope to converge upon the ideal recipe.

This approach can be described as *bottom-up*. That is to say, one does not prescribe in advance the tasks to be carried out. One merely submits to the machine a dataset to be ingested, dissected, and analysed. At the end of this indigestion, the artificial intelligence determines on its own the rules it must follow to achieve a result. Of course, it is up to the human to define the laws of learning, instructing the machine on how to read the data and how to use it to govern its behavior. The most striking illustration of this approach is that of a computer like HAL in *2001: A Space Odyssey*. Beyond being a true conversational agent, this machine is capable of determining its own operating procedure. Unlike an old Airbus onboard computer that endlessly recites the same protocols, it proves capable of adapting to new contexts, identifying its allies and enemies, and, if necessary, making sacrifices.

What is striking about these two paradigms is that they attempt to approximate human intelligence by different, albeit complementary, paths.

While the top-down strategy seeks to understand cognition by endowing it with aprioristic logico-deductive rules, the bottom-up strategy aims to imitate it through aposterioristic statistical associations. This division between pure reasoning and pure experience is not without echoing the division that Plato—and later Kant—drew between Sensibility and Reason. By proceeding logically and deductively, one is guaranteed never to produce a response that strays from the initial principles—but one is also doomed never to create anything new (the Babylonian recipe will never evolve). In the statistical approach, there is intrinsic creativity (one may discover new recipes), but one is never safe from an undesirable result (one may set the kitchen on fire). It is for this reason that I would like to propose, as a working hypothesis, that bottom-up statistical methods enable the imitation of synthetic a posteriori judgements, while top-down deductions reduce to analytic *a priori* judgements.

Hypothèse 3. *The top-down paradigm is capable only of producing analytic judgements, while the bottom-up paradigm is capable only of producing synthetic a posteriori judgements.*

It is important to stress that I am referring here to top-down and bottom-up programming paradigms—not to the theoretical paradigms of computer science, such as provability theory and computability theory. Although the syntax of axiomatization encouraged top-down programming strategies, there is nothing preventing it from being used to describe the operation of a typically bottom-up neural network. Statistical laws, after all, can be axiomatized; one can reason about these systems in a logico-deductive manner. Conversely, when I say that the semantics of the Turing machine partakes in the bottom-up programming strategy, this does not preclude the possibility of using a Turing machine to prove axioms. After all, top-down algorithms run just fine on a computer. Fundamentally, the Turing machine and axiomatic systems are two different representations of the same concept: that of an automatic computational process. They are thus not intrinsically top-down or bottom-up. Nevertheless, they have inspired contrasting ways of thinking about computer science—and of building its tools—that are diametrically opposed.

The chapters that follow aim to justify the hypothesis just laid out by describing, as precisely as possible, the technical tools that constitute each of these paradigms. The task will be to understand what in them is analytic or synthetic, but also to identify in advance certain shadowy zones. By critiquing these tools—by shedding light on the problems where they fail to deliver convincing judgements—I hope to carve out, with beak and claw, a space for a third kind of judgement, one that occupies the space between Sensibility and Reason. As we shall see, this judgement no longer takes quite the same form as Kantian synthetic *a priori* judgements.

Before launching into the analysis I have set myself, it may be helpful to pose a few preliminary questions: which of the two paradigms—top-down or bottom-up—is most widely used today? When exactly did they emerge? What kind of relationship do they entertain with one another? In light of the recent history of artificial intelligence, it seems plausible to say that we have witnessed a gradual shift from a predominantly top-down approach (to which this chapter is dedicated) to a growing inclination toward the bottom-up (which will be the subject of the next chapter).

Until the early 2000s, it was not uncommon to say that engineers, far from trying to design humanoid droids capable of speaking and behaving like human beings, were instead striving to develop algorithms dedicated to clearly defined tasks: playing chess, optimizing financial flows, managing logistics, etc. It was often repeated that computer science, when taken seriously, should dedicate itself to the development of weak artificial intelligence, not to the mythology of a thinking robot[4]. And yet today, we must admit that we have never been so close to designing the first truly general forms of artificial intelligence. The example of ChatGPT playing chess, answering questions in theoretical physics, or swapping, as if for sport, the laboratory scalpel for the poet's pen, has so captivated public discourse that it is now conceivable to work toward developing an algorithmic architecture that is context-independent.

This paradigm shift is all the more surprising in that, on the human side, we seem to be witnessing a phenomenon that is completely asymmetrical. For us citizens, the world of work appears to demand ever greater specialization. The days when universities trained universal scholars in the mold of Leonardo da Vinci are long gone. Today, it is a condition of existence for both workers and researchers to seize, by virtue of their competencies, a narrow patch of land within the vast and well-charted territories of knowledge. Thus we are in possession of tools whose origins (the data encompassed) and ends (the tasks they serve) increasingly elude the narrowness of our vision (hyperspecialization). Does this mean that we have no hope of understanding the machines we have brought into the world?

"Where danger grows," wrote Hölderlin, "grows also that which saves." It is from our collective intelligence, and from the vast sum of experiences and perspectives that compose our society, that self-adaptive machines have emerged—may this not also be where the solutions lie? Thanks to the Internet, that inexhaustible well of interconnected knowledge, big data are gathered daily. The Internet is what allows collective intelligence to be, so to

[4]To recall, weak artificial intelligence refers to software capable of performing specific tasks, such as playing chess, while strong artificial intelligence concerns a machine endowed with all human cognitive abilities, making it flexible, autonomous, and adaptable. The reader may refer to chapter for a definition of strong AI according to François Chollet. Note that some authors even add a third level, that of super artificial intelligence, which would possess consciousness and far exceed human capabilities.

speak, synthesized into our algorithms. Their behaviors, their biases, their reference systems are the statistical reflection of our opinions, our beliefs, our discourses, our actions. We are fashioning a world that, in every respect, mirrors our inner world. Yet we are doing so in such a way that it no longer seems to belong to us individually, but to humanity as a whole[5]. If so, then might it not also be through this collective intelligence that we will be able to continue understanding—and directing—the future developments of artificial intelligence?

To give historical and methodological structure to these reflections (which I touch on in chapters 16.2.2 and 16.2.2), I must now characterize the passage from the top-down to the bottom-up approach. A craftsman, before setting about the making of a piece of joinery, takes care to understand the tools he handles, and the impact these may have on his work. It is very likely that his relationship to construction is shaped by the arrival, in his workshop, of industrial machines; and once his chisel is replaced by automatic cutters, the principles and discipline he applies to achieve an optimal result are bound to change.

In the same spirit, the philosopher who reflects on artificial intelligence must understand how the evolution of the tool designed to model and understand human thought has, in turn, influenced the exercise of that very thought. That is why, in the pages that follow, I will briefly describe the shift from the top-down paradigm—well-defined and mastered—to the bottom-up paradigm. More particularly, I will be attentive to the epistemological principles, or implicit conceptions of the human mind, that are carried by each of these approaches.

6.1 The General Problem Solver

In 1967, within the walls of the Massachusetts Institute of Technology (MIT), a most unusual game of chess took place. One of the contenders, an eminent professor of philosophy named Hubert Dreyfus, faced off against the software Mac Hack VI. One can easily picture the scene. On one side, a patched-up white chair — the kind with a worn-out backrest, hunchbacked under the weight of years of study, of naps and feverish research, all passing through the university institute. Opposite it, a black screen streaked

[5]In this respect, it is noteworthy that the emergence of ChatGPT has deeply challenged the notion of intellectual property. It is as if the old regime of knowledge production—whereby an individual, by the power of their intellect, could imprint their own intelligence upon the external world—has been shaken by a new order. In this new regime, creations are collective and, so to speak, anonymous. When DALL·E, or any other automatic image generator, produces the portrait of an old woman, it assembles into a synthetic image the work of every photographer who has ever shared their creations online. In this wrinkled, faded face, aglow with the strength of generations, might there not be, somehow, a glimmer of your grandmother's smile?

with gleaming shapes. The human sits in a pose worthy of Rodin's *Thinker*. From the height of his chair, he challenges the imperturbable machine.

If not for the sizzling hum of the electronic circuits, one could hear a fly buzzing. Around the two opponents drifts a procession of pale faces, crooked glasses, tightly knotted ties. Brows furrow in perplexity. Sweat beads on bare temples. Lips release a faint smile of irony as the white queen, after a dazzling incursion, is caught off guard and captured.

A breath of relief passes through the part of the room occupied by the computer scientists. And they are not to be blamed: Hubert Dreyfus had previously had the audacity to declare that no computer would ever manage to defeat a ten-year-old child. The statement, to be sure, was provocative; but the boldness of this quip was rivaled only by another prophecy: "Within a decade," had they cried, backing the research sector with millions of dollars, "a computer will be able to defeat a chess champion!"

The philosopher Dreyfus was not the long-awaited champion. At most, he resembled a zealous amateur. A little clumsy. Brilliant, now and then. Perhaps too enthusiastic to be truly convincing. But he was no child either — neither prodigy nor beginner. Logically speaking, one might have expected a draw. That day, however, the machine prevailed.

Among the jubilant spectators stood a forty-year-old man. His forehead taller than the others, his glasses thicker, contrasted with the subtlety of his smile. His name was Herbert Simon. It was he who, ten years earlier, had asserted that the machine would beat the champion. He was aware, of course, of having been a bit too assertive, but the outcome of the match still offered, in favor of computational optimism, a fine swan song. "It was a magnificent game," he would later comment. "The suspense was intense between those two strategists, each with sparks of genius and devilish plans... with all the drama and disaster that so often accompanies such encounters[6]."

Far from being merely a match between man and machine, that MIT chess game symbolized, more fundamentally, the antinomic temptation of two opinions, two Muses, each with grandiose dreams. On one side stood the philosopher, deeply convinced that the machine, in its current state, could not imitate human thought in all its dimensions. On the other stood the computer scientist, imbued with the ambition to design a machine capable of solving all kinds of human problems. On the one hand, then, the Muse of flesh and clouds, with a face worthy of the most illustrious romantic paintings; on the other, the Muse of copper and silicon, resurrected straight from the film *Metropolis* (or from the hands of Frankenstein).

The confrontation on the chessboard, in itself, was nothing new in the history of artificial intelligence. Since the construction of the famous mechanical Turk, which once faced Napoleon, many such encounters had taken

[6]Translated from the *Oral History of Richard Greenblatt*, interviewed by Gardner Hendrie.

place between man and machine. Some of them, to be fair, were rigged — was there not a chess master hidden behind the well-oiled gears of the Turk? But what is truly inaugural in the game against Mac Hack VI is that, for the first time in human history, an operational description could be sketched of what a thinking machine might look like.

Toward the end of the 1950s, in fact, Herbert Simon had developed, alongside his colleagues Allen Newell and John Shaw, a piece of software entitled the *General Problem Solver* (GPS). Despite its rotund name, the program aimed to convince even the most skeptical of a computer's ability to solve any kind of problem. All one had to do was introduce, in the form of logical propositions, a description of the objects to be manipulated — the axioms of Euclidean geometry, the axioms of predicate logic, etc. — and the software would undertake, using operational research, a deduction of the best admissible solution. One of the program's specific features was that it separated the encoding of the problem-solving methods from the encoding of the problem itself. The authors of the paper summarize this mechanism using the diagram ??[7].

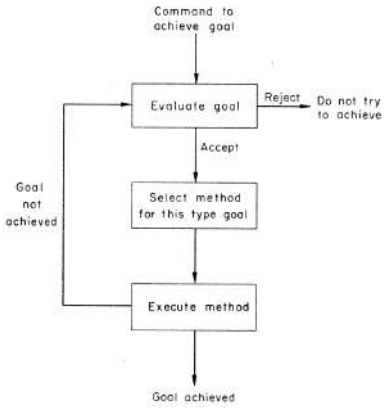


Fig. I—Executive organization of GPS

The fundamental principle of the *General Problem Solver* consists in taking as input the axioms of a theory and producing the maximum number of deductions. Due to computational complexity, the number of possibilities explodes rapidly, which necessitates refining the brute-force approach. One must determine, in particular, which data are relevant to consider upstream in the process, and which are interesting to produce downstream. To circumvent these computational difficulties, three key ingredients were introduced by Newell, Simon, and Shaw:

[7] A. Newell, J. C. Shaw, H. A. Simon, *Report on a general problem-solving program*, December 30, 1958.

-

- **Possibility Trees:** The machine methodically traverses a tree of possibilities. At the roots of this tree lie the available data (theory axioms and known relations). The branches, in turn, each represent a new deduction. This tree of possibilities must be traversed until one encounters the proposition to be proven (or its impossibility). This method is foundational to operational research, a major field in algorithmics.

- **Heuristics:** To avoid blindly traversing all possibilities, Newell and Simon made use of heuristics. These intuitive rules, pre-established by humans, are based on precise knowledge of the problem. An expert surgeon will not necessarily consider all hypotheses. His experience allows him to exclude, *a priori*, and often unconsciously, certain possibilities. In truth, conscious reasoning in humans operates only on a very small subset of cases, lost within the immensity of possibilities.

- **Lists:** To improve execution time, the information processed by the program is represented as lists of the type $A \rightarrow B \rightarrow C$, where the elements follow each other sequentially. This approach resembles the lambda calculus notation used by Church. It lies at the foundation of the LISP programming language, still in use today.

The machine is also equipped with two memory spaces: working memory and long-term memory. The former holds operations currently being calculated, while the latter contains all the facts known to the program, either because they are part of the initial premises, or because they have been demonstrated. This way of representing memory within the machine is not without evoking Plato's analogy of the wax tablet. It is as if the data were manipulable symbols, which can be imprinted somewhere within the machine's mechanisms, and recalled at will. To do so, it relies on two major deduction rules:

1.

2. **Forward chaining:** The program starts from the axioms and tries to combine them strategically to prove the formula under evaluation (in which case it is accepted), or to prove its negation (in which case it is rejected).

3. **Backward chaining:** The program starts from the formula and attempts to simplify it by reducing it to the theory's axioms (in which

case the formula is accepted) or to their negation (in which case it is rejected). It is enough for a single axiom to be contradicted for the formula to be disqualified.

Forward and backward chaining are like the two sides of an equality $A = B$. To prove this relation, you can either develop the right-hand side, or the left-hand side, or both successively. In Expert Systems, intelligence is conceived, purely and simply, as the manipulation of symbols in a digital memory space. Every object of knowledge must be introduced into the machine in the form of a number, so that it may then perform the calculation required to produce another number — the answer.

This approach, promising though it may be, is akin to a sophisticated calculator. It borrows its epistemological intuition from Hilbert's dream, when he declared that mathematics were nothing but the totality of tangible signs derivable, by a conscientious human being, on a blank sheet of paper. In the *General Problem Solver*, the symbols manipulated have no interpretation *a priori*. It is left to the one who inputs the data — and who gathers the machine's output — to endow them with meaning, declaring, for instance, that such a variable denotes a temperature, a coefficient of friction, and so on.

In the pages that follow, I will provide a more formal definition of the expert system. This will first require representing deductive reasoning, both forward and backward, through a mathematical language. Then, I will investigate how to solve, within that language, open-ended questions of the type "true" or "false."

6.2 Propositional Logic

To the extent that a machine manipulates symbols such as letters (A, B, C...), numbers (1, 2, 3...), or specific functional operators ($\rightarrow$, $\vee$...), it seems important to ask whether these symbols can truly represent human thought. If it were indeed shown that thought is something more than the mere manipulation of abstract signs, we would have in hand a decisive argument for drawing a boundary between the human being and the computer — at least, the computer as we know it today.

As it happens, this concern — that of representing ideas through signs — already preoccupied the ancient philosophers. Figures such as Aristotle and Chrysippus made significant contributions to developing and formalizing the study of reasoning, by showing which relationships and which sequences between ideas can, or cannot, legitimately be regarded as valid.

Logic consists, fundamentally, in an abstraction that reduces thought to a set of terms isolated from one another. In this sense, logic presupposes that cognitive processes can be flattened onto the page and represented in

linear form, each moment of thought necessarily leading to the next. This way of expressing the content of our reflections contrasts sharply with our subjective experience.

Logic, by imposing the linearity of reasoning, captures a fundamental aspect of the human mind: the relations of necessity between ideas. When we enter the domain of logic, we are no longer concerned with thought as it is lived from within, but with the objective result of thought. Certain associations between ideas seem spontaneously permitted, while others are not. It is, for example, acceptable to say that if Socrates is a man, and all men are mortal, then Socrates is mortal. On the other hand, it would seem suspicious to conclude from this that all men are Socrates. This is because there exists a relation of membership between the rational concepts we manipulate. Socrates belongs to the set of human beings, and those human beings in turn belong to the set of mortal beings.

Syllogisms do nothing more than express, in the most general form possible, the existence of this relation of membership that seems present in the human mind. By "general form", I mean an expression that holds true regardless of the terms involved. It must apply to Socrates and mortal men, but also to squares, rectangles, and quadrilaterals. This is why logic, beyond merely linearizing reasoning, also makes use of symbols. For instance, one may write:

$$\text{Si } A \rightarrow \text{B et si B } \rightarrow \text{C, alors A } \rightarrow \text{C}$$

In this expression, the arrow $\rightarrow$ designates implication and may be read as "if the term on the left is true, then the term on the right is also true", or alternatively, "whenever the term on the left is true, one also observes that the term on the right is true". It is easy to see that by substituting A with "being Socrates", B with "being a man", and C with "being mortal", one retrieves the syllogism from our earlier example: if being Socrates implies being a man, and being a man implies being mortal, then being Socrates implies being mortal. The reasoning remains valid even when we let $A =$ being a square, $B =$ being a rectangle, and $C =$ being a quadrilateral.

The expression above uses the vocabulary and grammar of what is called *propositional logic*. This logic, which is very simple to formalize, employs only a small number of ingredients, yet it can express a surprisingly wide range of thoughts. From a technical point of view, propositional logic can be defined as a language composed of an alphabet, a grammar, and a semantics, such that one can construct formulas that are deductively valid within an axiomatic system.

6.2.1 Alphabet

The alphabet is the set of all symbols used within the language. In propositional logic, we find only two types of symbols: an infinite, countable set

of propositional elements (denoted $p_1, p_2, ..., q_1, q_2, ..., r_1, r_2, ...$) as well as logical connectives ($\vee, \wedge, \neg, \rightarrow, \leftrightarrow$) and parentheses ("(" and ")"). We shall examine later what these symbols mean and how they are to be interpreted. For now, let us simply say:

- The negation $\neg$ is placed before a propositional symbol to express that what it denotes is false.
- The disjunction $\vee$ corresponds to an "or": it is placed between two propositional symbols to express that at least one of the two denotes something true. It does not exclude the possibility that both are true.
- The conjunction $\wedge$ corresponds to an "and": it is placed between two propositional symbols to express that both denote something true.
- The implication $\rightarrow$ is placed between two propositional symbols to express that if the first denotes something true, then the second does so as well.
- The bi-implication $\leftrightarrow$ is placed between two propositional symbols to express that both symbols denote either true things or false things. One says that they have the same *truth value*.

Note that one sometimes encounters non-standard operators, such as the Sheffer stroke ($|$) or the exclusive disjunction (noted $\oplus$ or w). However, it has been proven that these operators are redundant. A subset of them suffices to express all possible sentences[8].

6.2.2 Grammar

The grammar stipulates how sentences can be written using this vocabulary. In propositional logic, any sentence consisting of a sequence of propositional symbols such as $p_1, p_2, \ldots$ separated by logical operators can be written. The sentence $p_1 \rightarrow p_2$ is valid, while $p_1 p_2 \rightarrow p_3$ is not valid, because the symbols p_1 and p_2 are not separated by an operator. We do not know how to read them. Note that when several operators are involved in a single sentence, it is important to use parentheses to distinguish which operator applies to which symbol. This process is similar to what is done in arithmetic. The goal is to avoid any form of notational ambiguity.

If we denote by L the set of well-formed expressions of propositional logic, we can formalize what was just stated as follows:

L1 All propositional symbols are elements of L.

[8] For instance, one may work solely with $\vee$ and $\wedge$. Alternatively, one may use only the Sheffer stroke $|$.

L2 If $A \in L$ and $B \in L$, then

L2a $\neg A \in L$

L2b $A \vee B \in L$

L2c $A \wedge B \in L$

L2d $A \rightarrow B \in L$

L2e $A \leftrightarrow B \in L$

L3 Nothing else is an element of L.

6.2.3 Semantics

The sentences for which we have provided writing rules would be of no interest if they were not attached to a semantics. By semantics, I mean two things: (1) a certain way of reading the symbols, interpreting them, so as to match the sentence with reasoning in French, and (2) a way of determining whether the resulting sentence is true or false.

Point (1) requires associating the propositional symbols $p_1, p_2, \ldots, q_1, q_2, \ldots$ with propositions of the language. Point (2) consists in asking what the conditions are for a formula using these symbols, of the type $q_1 \rightarrow q_2$, to be itself true or false.

It is important to note that propositional logic is a bivalent logic. It states that only two truth values exist: true or false. This conviction that something is either true or false is known as the principle of the excluded middle. Aristotle is its defender when he writes:

> *It is not possible for there to be any intermediate between contradiction, but it is necessary to affirm or deny one thing of one thing*[9].

He further clarifies that something is true or false only in thought, meaning in the way an idea is either associated or not with another:

> *Neither the false nor the true are in things, as, for example, if the good were the true, and the bad, the false. They exist only in thought; moreover, simple notions, the conception of pure essences, produce nothing of the kind in thought*[10].

By convention, we associate each propositional element $q_1, q_2, ...$ with a truth value depending on the interpretation of that term. The truth value is 1 if the propositional element is true, and 0 if it is false.

[9] Aristotle, *Metaphysics*, IV, 7, 1011b23-24.
[10] Aristotle, *Metaphysics*, III, 25.

As an example, let p_1 = "The person reading this sentence is a human being" and p_2 = "The person reading this sentence is mortal." These propositions can be true or false depending on the context. When you read this sentence, we have p_1 = true and p_2 = false, but if a god from mythology were to read it, we would have p_1 = false and p_2 = true.

We then use rules to determine if a sentence such as $q_1 \wedge q_2$ should be judged true or false. These rules are determined by following the intuitive meaning of the logical operators: since the symbol $\wedge$ represents a logical "and", the sentence $q_1 \wedge q_2$ will be true if and only if both q_1 and q_2 are true; the sentence $q_1 \vee q_2$, on the other hand, is true if and only if one of the two propositional elements is true, etc.

These rules are summarized in the table below, also called the *truth table*:

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

In the example given earlier for q_1 and q_2, the sentence $q_1 \rightarrow q_2$ reads "If the person reading this sentence is called Socrates, then the person reading this sentence is male." Intuitively, this sentence can only be false if the person reading it is a human being and that person is immortal. In our world, this is not possible, but that does not mean that the sentence is true for all eternity, and this for two reasons.

The first is that one can conceive of worlds—mythological ones, admittedly—where a human being is immortal. In Hindu mythology, the Chiranjivis are human beings who have been granted the gift of immortality. The best known among them is Hanuman.

The second reason is that, even while remaining within our known world, it is possible to give the propositions q_1 and q_2 a different interpretation from that of our example. One could say q_1 = "the person reading this sentence is a human being" and q_2 = "the person reading this sentence is male." Under this new interpretation, it is easy to find cases where the formula $q_1 \rightarrow q_2$ is false. It is enough to imagine that a woman reads this sentence.

An interesting question in logic is whether there are sentences whose truth value does not depend on the interpretation of the elements from which it is composed. We observe three scenarios:

- **Tautologies:** Some sentences are always valid, regardless of the interpretation of the propositional elements. For example, the sentence $q_1 \vee \neg q_1$ is a tautology that says: "either q_1 is true, or q_2 is false." It is

a logical law.

- **Contradictions:** Some sentences are never valid, regardless of the interpretation of the propositional elements. For example, the sentence $q_1 \wedge \neg q_1$ is a contradiction that essentially expresses: "either q_1 is true and false at the same time."

- **Contingencies:** Some sentences can be true or false, depending on the interpretation of the terms. For example, $q_1 \wedge q_2$ is true if and only if q_1 and q_2 are true. This can be obtained, for instance, by taking q_1 = "the sky is blue" and q_2 = "the sun is a star."

Tautologies are particularly interesting, as they can be used to perform reasoning. The fact that they are never false allows us, at times, to simplify certain expressions or to show that two relations that we thought were very different are, in reality, equivalent.

6.2.4 Formulas

It is important to note that it is possible to write logical propositions where the variables are not the propositional elements $p_1, p_2, ...$, but rather formulas $A, B, ...$ What, exactly, is the difference between $p_1 \vee p_2$ and $A \vee B$? In the first case, the sentence must be interpreted as "either we have p_1, or we have p_2", where p_1 and p_2 can represent any idea. They could be associated with the semantics p_1 = "red" and p_2 = "black", just as they could be associated with the semantics p_1 = "small" and p_2 = "large".

The letters A and B in the second case, however, represent any propositional formula. Thus, we can substitute the expression $p_1 \rightarrow p_2$ for the letter A and $p_1 \vee (p_2 \vee p_3)$ for the letter B. In the specific case where A equals p_1 and B equals p_2, we find the expression $p_1 \vee p_2$, but this is a special case. In other words, the use of the letters A and B is much more general than that of propositional symbols. While propositional symbols are used to express relationships uniting ideas, any arbitrary individual elements of thought, the capital letters, on the other hand, are mobilized to express relationships uniting reasoning processes with one another.

6.2.5 Axiomes

Human reasoning is characterized, among other properties, by the ability to make deductions. These involve starting from certain logical propositions and assembling them to deduce more general truths. The typical example of deduction is found in Aristotle's syllogistics: if we know that Socrates is a man ($s \rightarrow H$) and that men are mortal ($H \rightarrow M$), we deduce that Socrates

is mortal ($S \to M$). In modern terms, we would speak of transitivity. This mental process highlights the existence of logical laws that dictate which deductions are valid and which are fallacious.

In propositional logic, these laws can be written as a set of axioms. These axioms take as variables the formulas A, B, C of the propositional language and tell us how they can be associated with one another. In other words, axioms are sentences that are always true, no matter how the terms A, B, and C are interpreted.

$AS1$	$A \to (B \to A)$
$AS2$	$(A \to (B \to C)) \to ((A \to B) \to (A \to C))$
$AS3$	$((A \to B) \to A) \to A$
$AS4$	$(A \wedge B) \to A$
$AS5$	$(A \wedge B) \to B$
$AS6$	$A \to (B \to (A \wedge B))$
$AS7$	$A \to (A \vee B)$
$AS8$	$B \to (A \vee B)$
$AS9$	$(A \to C) \to ((B \to C) \to ((A \vee B) \to C))$
$AS10$	$(A \leftrightarrow B) \to (A \to B)$
$AS11$	$(A \leftrightarrow B) \to (B \to A)$
$AS12$	$(A \to B) \to ((B \to A) \to (A \leftrightarrow B))$
$AS13$	$\neg\neg A \to A$
$AS14$	$(A \to B) \to ((A \to \neg B) \to \neg A)$

Although these axioms may seem surprising at first glance, they correspond to quite familiar reasoning methods. Axiom 4, for example, tells us that if we know that $A \wedge B$ is true, then we can deduce that A is true. This is written as:

$$A \wedge B \vdash A$$

The symbol $\vdash$ used in this expression indicates that there exists a logical deduction from a statement called the premise $(A \wedge B)$ to write a new statement called the conclusion (A)[11]. These deductions are defined as a set of steps, or lines of symbols, that rely on known true statements to obtain new true statements.

Note that the axioms I presented above are also equipped with an ad hoc rule of inference called Modus Ponens . This rule states that if we have A and $A \to B$, then we can write B. This rule is written as:

$$A, A \to B \vdash B$$

[11] If there are multiple statements to the right of the symbol $\vdash$, at least one of them must be true. However, this case is rarely encountered and constitutes a kind of convention.

We can now give a formal definition of the concept of logical deduction as a reasoning process consisting of a number of steps, represented as lines, and justified by the axioms.

Proposition 2. *An axiomatic deduction of B from the premises $A_1, A_2, ..., A_n$ is a list of lines such that each line is:*

- *a premise, or*
- *an axiom of classical propositional logic (AS1–AS14), or*
- *a result D of the MODUS PONENS rule: if C is on line i in the list and $C \rightarrow D$ is on line j of the list, D can be derived on line k (where $k > i$ and $k > j$) of the list*

and where B is the final element of the list.

Note that I used the term axiomatic deduction here, meaning that the deduction involves only the axioms. In the practice of an expert system, it is convenient to use, at certain stages of reasoning, deductions that have already been proven earlier.

To illustrate this approach, let us first assume that we are able to prove Modus Tollens axiomatically. Modus Tollens, sometimes called the contrapositive, is a well-known relation in logic, which states that if A implies B and B is false, then A cannot be true. If A were true, we would have B. In logic, this is written as:

$$\neg B, A \rightarrow B \vdash \neg A$$

An axiomatic proof can be given using axioms AS1 and AS14:

1	$\neg B$	PREMISE
2	$A \rightarrow B$	PREMISE
3	$\neg B \rightarrow (A \rightarrow \neg B)$	AS1
4	$A \rightarrow \neg B$	MODUS PONENS (lines 1 and 3)
5	$(A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$	AS14
6	$(A \rightarrow \neg B) \rightarrow \neg A$	MODUS PONENS (lines 2 and 5)
7	$\neg A$	MODUS PONENS (lines 4 and 6)

It is important to note that the symbols in axiom AS1, found on line 3, are not identical to the symbols I used when introducing this axiom. Axiom AS1 was indeed defined as:

$$A \rightarrow (B \rightarrow A)$$

However, in the proof, I used:

$$\neg B \rightarrow (A \rightarrow \neg B)$$

To understand this substitution, we must remember that the symbols A and B appearing in the writing of axiom AS1 represent any formula. It is therefore perfectly permissible to choose the formula $\neg B$ in place of A and the formula A in place of B. A similar process applies in the case of axiom AS14, which was used on line 5.

Now that the axiomatic proof of Modus Tollens is complete, it is convenient to use Modus Tollens within a deduction without having to go through the axioms again. For example, if I am making a deduction where, in previous steps, I have already established that $\neg B$ and $A \rightarrow B$, I can immediately write $\neg A$ on the next line of my reasoning. From what I have just said, a new definition follows.

Proposition 3. *A deduction (not necessarily axiomatic) of B from the premises $A_1, A_2, ..., A_n$ is a list of lines such that each line is:*

- *a line authorized by axiomatic deduction, or*
- *the result D of a rule $C_1, C_2, .., C_n \vdash D$, where the premises $C_1, C_2, .., C_n$ all appear among the previous lines in the list*

and where B is the final element of the list.

6.3 Expert Systems

The propositional logic that has been presented can be used by an Expert System to make deductions. Such a system then has three types of components:

1. **The axioms of logic:** the fundamental laws of human reasoning, represented in the simplest and most abstract form.

2. **The laws of the theory:** the fundamental laws of the disciplinary field within which the Expert System operates. If it is a machine designed to diagnose cancerous tumors, it will need to be equipped with a certain number of hypotheses about how clinical observations (EEGs, X-rays, etc.) can be used to establish valid conclusions (presence of a benign tumor, malignant tumor, etc.). These hypotheses are written in the form of logical laws and added to the axioms of logic itself.

3. **The data of the problem:** The clinical observations available about the patient must be encoded in the Expert System as propositional elements of types $q_1, q_2, \ldots$, to which truth values are assigned. If q_1 represents the presence of a lump and there is no lump, the machine will be given $q_2 = \text{false}$. If there is a lump, $q_3 = \text{true}$. If uncertain, the formula $q_1 \vee q_2$ will be added to the machine.

It will be understood that it is crucial for the logical language used to be sufficiently precise to encode both the laws of the theory and the problem data. However, propositional language has certain shortcomings. First, it does not allow for expressing truth values other than true or false. Although this distinction may seem sufficient at first glance, it fails to differentiate between what is "probably false" and what is "certainly false." One might therefore be tempted to assign truth values not strictly equal to 1 or 0, but instead within the interval $[0; 1]$. Such an approach leads to a probabilistic logic.

Another problem arises because propositional logic does not support contradictions. If the system contains two paradoxical premises, such as p_1 and $\neg p_1$, it becomes possible to prove both something and its opposite. Indeed, one could rely on p_1 to prove certain results, and then on $\neg p_1$ to prove the opposite result. This property of the system is called the principle of explosion and is written as:

$$A, \neg A \vdash B$$

An axiomatic proof can be given as follows:

1	A	PREMISE
2	$A \to (\neg B \to A)$	AS1
3	$\neg B \to A$	MODUS P. (lines 1 and 2)
4	$\neg A$	PREMISE
5	$\neg A \to (\neg B \to \neg A)$	AS1
6	$\neg B \to \neg A$	MODUS P. (lines 1 and 2)
7	$(\neg B \to A) \to ((\neg B \to \neg A) \to \neg\neg B)$	AS14
8	$(\neg B \to \neg A) \to \neg\neg B$	MODUS P. (lines 3 and 7)
9	$\neg\neg B$	MODUS P. (lines 6 and 9)
10	$\neg\neg B \to B$	AS13
11	B	MODUS P. (lines 9 and 10)

The principle of explosion is particularly troublesome for expert systems, as the databases used to enable the machine to reason may contain contradictions. For instance, consider a system designed to diagnose the presence of a cancerous tumor. It may be useful to provide this machine with all the clinical observations established about a patient. These observations could

have been made at different times by different people. They depend on measurement devices, the specialist's skills, their prior assumptions, and may also contain errors. Therefore, it is likely that contradictions will appear between the data. However, we would want the expert system, if it concludes that a tumor is present, not to do so based on the principle of explosion, otherwise, we could observe absurd reasoning such as: "The patient has a fever of 37 °C and 36°C. This is a contradiction. Therefore, I deduce that they have cancer."

One limitation of propositional logic, in the context we are interested in, is the inability to distinguish between possibility and necessity. We would like to be able to distinguish between sentences like "the appearance of such a symptom makes the existence of a tumor impossible" and "the appearance of such a symptom makes the existence of a tumor possible." However, in propositional logic, the implication $\rightarrow$ always denotes a relationship of necessity, without exception. To address this, propositional logic must be refined by adding the symbol $\Box$ in front of properties that are necessary and the symbol $\Diamond$ in front of those that are merely possible.

A final issue with propositional logic is its expressive poverty. It does not allow for distinctions between the subject and the predicate using propositional elements. For each subject i of interest within a set of objects and for each property j within a set of properties, one must create a term p_{ij} that is true if subject i possesses property j. This language, however, does not allow the construction of complex relationships between properties and their variables. It also cannot accommodate functions that transform variables, such as mathematical operations like addition, subtraction, or multiplication. Thus, it is impossible to ask questions like: "Is the product of two arbitrary numbers divisible by two?" or "Is there a number greater than the sum of two arbitrary numbers?" To introduce this degree of refinement, it may be helpful to turn to predicate logic, which introduces variables, predicates, functions, and operators such as "for all" ($\forall$) and "there exists" ($\exists$).

In the remainder of this chapter, despite the shortcomings I have outlined, I will focus solely on propositional logic. The reason for this is that the general conclusions drawn from propositional logic regarding expert systems can be transposed to more advanced logics. My goal here is not to exhaustively cover all that is done in the field of algorithms, but rather to introduce the reader to the essential concepts of the Top-Down approach. The reader will also notice that the tools discussed in this state of the art will be used or addressed in later parts of the thesis.

6.4 Conjunctive Normal Form

When children play "neither yes nor no", they impose restrictions on the words they can use to express their thoughts. However, this does not dimin-

ish their expressive power. Everything that language allows us to convey using the words "yes" and "no" can also be expressed with synonyms (though certain emotional nuances may be lost). In a similar vein, we might wonder: what are the smallest sets of operators $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$ needed to express the laws of propositional logic?

This is an important question for the development of Expert Systems. If it were established that there were many different ways to express the same logical laws, this would complicate machine deductions. The machine would, in fact, need more time to determine whether two expressions are identical. Moreover, it might produce relations that are merely redundant rewritings of the information we already possess[12]. That is why it is worth asking whether there exists a unified way to express logical relations? Ideally, this unification must be done so as to satisfy certain criteria:

- **Simplicity:** This unified form must minimize the number of logical operators;
- **Expressive power:** It must not restrict the machine's expressiveness;
- **Computability:** It must maximize the speed at which the machine performs calculations;
- **Elegance:** It must be elegant enough for a human being to understand the final result (and ideally the intermediate steps of reasoning).
- **Traceability:** It must be clear enough for a human to understand the intermediate steps of reasoning.

To embark on this path, it may be helpful to observe that the implication $A \rightarrow B$ is equivalent to $\neg A \vee B$. A simple inspection of the truth table confirms that the implication $A \rightarrow B$ is true if and only if the antecedent A is false, or the consequent B is true.

This property is not at all intuitive, for it means that if the consequent is trivially true, one can choose absolutely anything as the antecedent. For instance, I could take as an antecedent, "If it is raining and it is not raining" (a contradiction), and construct the sentence P: "If it is raining and it is not

[12] In general, propositional logic allows only the deduction of tautologies from one another, so whatever is obtained will always bear the mark of triviality. However, there are cases in which logical deduction can help human beings clarify certain problems. It is, nevertheless, reasonable to think that if Simon and Newell had remembered Kant's teachings, they would not have asserted so confidently that their General Problem Solver—a machine specialized solely in the resolution of analytical judgements—could answer all possible questions.

raining, then either I am alive or I am not alive." This sentence P will be true, and for two reasons.

The first reason is that the antecedent, "If it is raining and it is not raining," is always false. Now, an implication can only be false when something true leads to something false. The implication merely asks whether an event or state of affairs A can be accompanied by an event or state of affairs B. If A never occurs, nothing false can be said of it.

The second reason is that the consequent—"either I am alive or I am not alive"—as found in proposition P, is a tautology, always true. It is therefore impossible to falsify proposition P. The situation is akin to a smoke that continuously hangs, in all times and all places, above the surface of the earth. Since it is ever-present, this smoke could just as well be said to accompany fire as to accompany water, air, ice, or wind. Thus it would be as accurate to say, "there is no ice without smoke," as it is to say, "there is no fire without smoke."

These explanations I have provided show that the implication $\rightarrow$ is purely formal: it does not account for any substantial connexion between the ideas being expressed. In our example, there is no logical connexion between "it may rain and it does not rain" and "either I am alive or I am dead." The only thing that matters is to know in which case one of these propositions can be true while the other is false, and vice versa. It is probably not inappropriate to trace this formal view of things back to Wittgenstein, for whom the world is composed not of subjects and properties, but of atomic facts that are connected to one another.

> *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.*

It is due to the absence of material connexions expressed between ideas that the implication $\rightarrow$ in propositional logic can be replaced by the symbols $\vee$ and $\wedge$ without weakening the expressiveness of the language. I will show later that this simplification will speed up the efficiency of calculations. Furthermore, it does not seem any less elegant when read by a human mind. It is even simpler, at first glance, to express oneself solely with "and" and "or" than with "if... then."

Just as $\rightarrow$ can be eliminated, it is also possible to replace $A \leftrightarrow B$ with $A \rightarrow B \wedge B \rightarrow A$, leading to $(\neg A \vee B) \wedge (\neg B \vee A)$. This latter expression is what is called a conjunctive normal form, defined as a conjunction ($\wedge$) of disjunctions ($\vee$). From a technical point of view, it is possible to show that any logical proposition can be reduced to a conjunctive normal form. To do so, it is necessary to introduce some basic vocabulary.

Définition 7. *A literal l is a propositional term preceded, or not, by the symbol $\neg$.*

Définition 8. *A clause in propositional logic is a countable set of n conjunctions of the form $l_1 \vee l_2 \vee ...l_n$, where $n \in \mathbb{N}$ and $l_1, l_2, ..., l_n$ are literals.*

Définition 9. *A conjunctive normal form is a countable conjunction of n clauses of the form $C_1 \wedge C_2 \wedge ...C_n$, where $n \in \mathbb{N}$ and $C_1, C_2, ..., C_n$ are clauses.*

Proposition 4. *Every formula in propositional logic can be written in conjunctive normal form.*

Proof. □

When a formula is written in conjunctive normal form, it will be sufficient, to know whether it is true, to verify that all the clauses $C_1 \wedge C_2 \wedge ...C_n$ it consists of are true. Thus, for each clause $l_1 \vee l_2 \vee ... \vee l_n$, we need to verify whether (1) there is a deduction from this clause to the laws and axioms of the theory (descending method), or whether (2) there is a deduction from the laws and axioms of the theory to this clause (ascending method).

To verify that a formula written in conjunctive normal form $C_1 \wedge C_2 \wedge ...C_n$ is false, one must check that at least one of the clauses $C_1, C_2, ..., C_n$ is false. This amounts to determining whether the negation of one of these clauses, denoted $\neg C_1 \vee \neg C_2, ..., \neg C_n$, is true. The negation of a clause $l_1 \vee l_2 \vee ...l_n$ (where $l_1, l_2, ...l_n$ are literals) is written as $\neg(l_1 \vee l_2 \vee ... \vee l_n)$, and is equivalent to $(\neg l_1 \wedge \neg l_2 \wedge ... \wedge \neg l_n)$. This transformation follows from De Morgan's laws (see the figure). Verifying the negation of a clause thus amounts to verifying the truth of a conjunctive normal form where each clause trivially contains only one term.

From these analyses, it follows that, to study the truth or falsity of a conjunctive normal form formula $C_1 \wedge C_2 \wedge ...C_n$, the expert system must always, ultimately, attempt to reduce the clauses $C_1, C_2, ..., C_n$ (or their negations) to the laws of the theory and the axioms. Search algorithms can, therefore, conveniently focus on how to manipulate clauses, transform them, and combine them using the laws of propositional logic. It is possible to design very efficient computational methods by following this strategy.

6.5 Satisfiability Problem (SAT)

The purpose of an expert system is, in essence, to be capable of answering any question posed to it. This question, as we have seen, must be introduced into the machine using a logical language, and this logical language, ultimately, can be presented in conjunctive normal form. It follows that the expert system must, ideally, be able to determine the truth or falsity of an expression of the type:

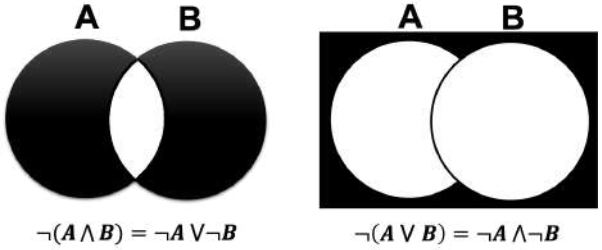


Figure 6.1: De Morgan's laws allow us to express conjunctions as negations of disjunctions and disjunctions as negations of conjunctions.

$$\wedge_i^m \vee_i l_{m,i}$$

where n is the number of clauses, m_i is the number of literals contained in the i-th clause (with $0 \geq i \geq n$), and where the $l_{m,i}$ are literals.

We have seen that a logical formula is either tautological, contradictory, or contingent. In the previous section, I indicated how the expert system could validate tautological propositions and invalidate contradictory propositions, by analysing the clauses one by one and attempting to relate them to the laws of the theory and the axioms of logic. However, it remains to be asked what happens in the presence of contingent propositions?

A formula F is contingent if and only if there exist situations in which it is true and others in which it is false. In the language of propositional logic, this means that it is possible to assign a truth value to the propositional terms $p_1, p_2, ...$ of which the formula F is composed in such a way that F is true or such that F is false.

A large number of questions that an expert may face will not immediately receive a conclusive answer. It is even likely that the more qualified the expert, the less quick they will be to provide a definitive answer. Where the novice traces a direct path, often littered with prejudices, between observation and conclusion, all kinds of hypotheses will present themselves to the mind of a more experienced individual. The doctor trying to determine if their patient has cancer will start by formulating contingent answers. They will not say, "Yes, the data allow us to conclude," but will respond, "If this hypothesis is confirmed, then I can conclude that the patient has cancer," and proceed with further observations.

This is why it is desirable that the Expert System, just like a human being, should be able to produce contingent propositions. Faced with a difficult question, for which it cannot say whether it is true or false given the available data, the system must be able to exhibit the conditions that will make this proposition true in order to guide future research. This problem

consists precisely in identifying which propositional terms must be true, or more generally, what relationships must exist between these propositional terms, for the formula under consideration to sometimes be true and sometimes false.

The problem of finding the conditions that satisfy a conjunctive normal form formula $C_1 \wedge C_2 \wedge ... \wedge C_n$ splits into two typical cases. In the particular case where the clauses $C_1, C_2, ..., C_n$ each contain only one or two terms, the problem is called 2-SAT (where SAT is an abbreviation of the word "satisfiability"). In the case where a clause contains at most three terms, we speak of the 3-SAT problem. In the general case, where no restrictions are imposed on the number of literals, we speak of CNF-SAT (CNF being an acronym for *Conjunctive Normal Form*). I will describe these problems and the ways to solve them.

6.5.1 Complexity of 2-SAT

The 2-SAT problem, relatively simple, can be solved in linear time. By this, I mean that the number of resolution steps required to determine whether 2-SAT is satisfiable grows linearly with the size of the problem. By "size of the problem," I refer to the total number of clauses $C_1, C_2, ..., C_n$ involved in the formula to be analysed. By "resolution steps," I refer to any operation required by an algorithm, such as: reading a symbol, writing a symbol, etc.

In other words, there exists a linear function between the number of steps required for the resolution algorithm and n, the number of clauses in the formula $C_1 \wedge C_2 \wedge ... \wedge C_n$. In mathematical notation, we say that the problem has a time complexity of $\mathcal{O}(n)$, where the stylized letter $\mathcal{O}$ denotes any linear function taking n as its argument.

I will provide here a sketch of the proof. First of all, it is important to note that each clause $l_1 \vee l_2$ in the 2-SAT problem can be rewritten using an implication in two different ways: $\neg l_1 \rightarrow l_2$ or $\neg l_2 \rightarrow l_1$. For example, consider the following formula:

$$(p_1 \vee p_2) \wedge (\neg p_2 \vee p_3) \wedge (\neg p_3 \vee p_1) \wedge (\neg p_1 \vee p_2)$$

From these four clauses, we can write eight implications. I will list them below after eliminating any double negations ($\neg\neg p_i$ simplifies to p_i):

- $\neg p_1 \rightarrow p_2$
- $\neg p_2 \rightarrow p_1$
- $p_2 \rightarrow p_3$
- $\neg p_3 \rightarrow \neg p_2$

- $p_3 \rightarrow p_1$
- $\neg p_1 \rightarrow \neg p_3$
- $p_1 \rightarrow \neg p_2$
- $\neg p_2 \rightarrow \neg p_1$

By combining formulas (1), (3), and (5), we can derive the following chain of implications:

$$(\neg p_1 \rightarrow p_2) \wedge (p_2 \rightarrow p_3) \wedge (p_3 \rightarrow p_1)$$

This chain indicates that if p_1 is false, then p_2 is true (first clause). But if p_2 is true, then p_3 must be true (second clause). Finally, if p_3 is true, then p_1 is true. It follows that if p_1 is false, then p_1 must be true.

We therefore conclude that p_1 must be true. Still, it remains to ask whether the fact that p_1 is true might, by a similar line of reasoning, lead us to conclude that p_1 must be false. Indeed, we find that by combining formulas (7) and (8), we can derive the following chain:

$(p_1 \rightarrow \neg p_2) \wedge (\neg p_2 \rightarrow \neg p_1)$

We thus conclude that p_1 cannot be true either. It follows that the formula is unsatisfiable.

This method illustrates a strategy for solving the 2-SAT problem. First, we must write all the $2n$ implications associated with the n clauses. Then, we must check for the existence of a chain going from a variable to its negation and from its negation back to the original variable.

The naive approach illustrated above consists, for each variable p_i, in first checking whether there exists a chain leading from p_i to $\neg p_i$, and then from $\neg p_i$ to p_i. To find a chain between two variables, we may require up to $\mathcal{O}(n^2)$ steps[13]. If we do this for every variable, the complexity rises to $\mathcal{O}(n^3)$. This is why the crux of the problem lies in structuring the information so that it is not necessary to traverse the list of implications more than k times, where k is a constant independent of n.

A solution was suggested in 1979 by Aspvall, Plass, and Tarjan[14]. It consists of representing the connexions between propositional terms using a directed graph. This graph is a mathematical structure composed of points, also called nodes, some of which are connected by directed arrows. We ensure that:

[13] A chain may involve, in the worst case, nearly $2n$ implications. Each implication requires traversing the list of $2n$ implications. The total complexity is therefore on the order of $2n \times 2n$, or $\mathcal{O}(n^2)$.

[14] Bengt Aspvall, Michael F. Plass, and Robert E. Tarjan, *A linear-time algorithm for testing the truth of certain quantified boolean formulas*, Information Processing Letters, vol. 8, no. 3, 1979, pp. 121-123.

- A point is created for each propositional term p_i in the original formula;
- A point is created for the negation $\neg p_i$ of each propositional term in the original formula;
- An arrow connecting points x and y is created if and only if there exists an implication $x \rightarrow y$ (where x and y each represent a propositional term or its negation).

This graph is represented in the computer system by creating, for each point x in the graph, a sequence of variables $< x, y_1, y_2, ..., y_l >$. The variable x represents the propositional term associated with this point in the graph. The variables $y_1, y_2, ..., y_l$ are the set of points connected to x by an outgoing arrow.

The problem is then to create this graph in $\mathcal{O}(n)$ operations and to determine if there exists a cycle that allows traversal from a given variable to its negation, and then from this negation back to the original variable. This can be done using either Kosaraju's algorithm, which traverses the graph's nodes twice, or Tarjan's algorithm, which traverses them only once. I will not describe these methods, but they involve cleverly organizing the information stored at each point.

6.5.2 Complexité de 3-SAT

The 3-SAT problem presents clauses that each contain at most three literals $l_1 \vee l_2 \vee l_3$. The method described in the case of 2-SAT is no longer valid because it is no longer so obvious to represent this clause solely through implications. We should indeed begin by writing six implications:

- $\neg l_1 \rightarrow (l_2 \vee l_3)$
- $\neg(l_2 \vee l_3) \rightarrow l_1$
- $\neg l_2 \rightarrow (l_1 \vee l_3)$
- $\neg(l_1 \vee l_3) \rightarrow l_2$
- $\neg l_3 \rightarrow (l_1 \vee l_2)$
- $\neg(l_1 \vee l_2) \rightarrow l_3$

Each disjunction $l_2 \vee l_3$ appearing in these equations must be replaced by $\neg l_2 \rightarrow l_3$ or $\neg l_3 \rightarrow l_2$. This can lead to some difficulties. First of all, one

might encounter expressions containing the negation of an implication, such as:

$$\neg(\neg l_2 \rightarrow l_3) \rightarrow l_1$$

It is not necessarily easy to manipulate the negation of an implication in the graph representation given earlier. Moreover, and this is the second difficulty, the implications we are dealing with are chains of three literals, like:

$$\neg l_1 \rightarrow (l_2 \rightarrow l_3)$$

The question of whether it is possible to adapt this new form of the problem to the Tarjan or Kosaraju algorithm, or any other method that allows a solution with complexity $\mathcal{O}(n)$, or even $\mathcal{O}(n^k)$, where k is any positive integer, remains an open problem, and one that the Clay Institute has awarded a million-dollar prize for. The currently known solutions are of the form $\mathcal{O}(2^n)$. All we can say with certainty about 3-SAT is that, although it is often difficult in practice to find a solution (the solving method), it is rather obvious to verify whether a given answer s is a satisfactory solution to the problem (the verification method).

This comes down to proving that there exists a polynomial-time verification method for the problem. If a certain value is given to the propositional terms $p_1, p_2, ...$ appearing in the formula, it is enough to use the truth table to determine whether the formula as a whole is true or false. This verification requires a number of steps proportional to the number of clauses in the formula (recall that this formula is written in conjunctive normal form); in other words, it is enough to scan the formula from left to right once. We conclude that 3-SAT thus has a complexity of $\mathcal{O}(n^k)$, where n is the number of symbols in the formula and k is any integer. This means that 3-SAT is of polynomial complexity.

From a terminological point of view, all problems such as 3-SAT, for which the verification method has linear-time complexity, belong to the NP class of problems. The acronym NP stands for Non-deterministic Polynomial, referring to the fact that a non-deterministic Turing machine would take polynomial time to solve it. The exact definition between deterministic and non-deterministic machines will be addressed later in this thesis. Suffice it to know for now, and to keep it brief, that the deterministic machine explores only one path at a time, while the non-deterministic machine explores several paths simultaneously. In other words, a deterministic machine can try all the answers and see if they work, while a non-deterministic machine must figure out which is the correct answer.

Définition 10. *An NP problem is a problem for which a given solution can be verified (rejected or accepted) in polynomial time.*

6.5.3 Complexity of CNF-SAT

One might wonder what the complexity is for problems like 4-SAT, 5-SAT, ... k-SAT, where k is any positive integer. Indeed, since the transition from 2-SAT to 3-SAT radically changes the nature of the problem, it is reasonable to think that the same would happen when moving from 3-SAT to 4-SAT. However, it turns out that this is not the case. It is possible to show that the k-SAT problem (with $k > 2$) can be reduced to the 3-SAT problem, in the precise sense that if a resolution method with a given $\mathcal{O}$ complexity is known for 3-SAT, this same method will work for 4-SAT without changing the complexity order.

This property is what allows all problems of the 3-SAT or higher type to be grouped under the term CNF-SAT. In other words, determining the satisfiability of a formula in conjunctive normal form, except in cases where these clauses contain at most 2 literals per clause, is to solve one and the same general problem.

Schaefer's Dichotomy Theorem generalizes this result by showing the precise conditions under which a variant of the SAT problem can be assimilated, in terms of complexity, either to the 3-SAT problem or to the 2-SAT problem.

6.5.4 Spatial Complexity

In addition to the time complexity that we have just analysed, which seeks to determine the number of steps required in the worst-case scenario, it is important to keep in mind that an algorithm also uses memory resources. It keeps track of a certain number of symbols during its execution. If this number becomes too large, the storage space may overflow before the algorithm has converged on a solution. This is why it is essential to determine the relationship between the size of the problem and the number of symbols that must be stored in memory at any given moment during the process, in the worst-case scenario.

We say that a program has a spatial complexity of $\mathcal{O}(n)$ if the number of symbols stored in memory in the worst case is proportional to the size of the problem. Here, the size of the problem is given by the number of propositional elements $p_1, p_2, ..., p_n$. In 1970, Savitch proved a theorem[15] that implies that 2-SAT has spatial complexity $\mathcal{O}(\log(n))$. This means that the number of symbols to be stored is a logarithmic function of n. This is one of the few results that allows us to say that a problem with a determined time complexity also has a spatial complexity. In fact, the spatial complexity of 3-SAT is not known.

[15] Savitch's theorem, stated as simply as possible, asserts that if a problem has spatial complexity $\mathcal{O}(\log(n))$ with a deterministic Turing machine, it also has spatial complexity $\mathcal{O}(\log^2(n))$ for a non-deterministic machine.

6.6 NP-Complete Problems

At this point, I would like to anticipate an objection in the reader's mind. Although I have shown that k-SAT problems (with $k > 2$) are difficult to solve, the question remains whether they are of practical interest. Indeed, it is possible that most of the questions we ask Expert Systems could be represented in conjunctive normal form with, at most, two literals per clause. If that were the case, there would be nothing troubling about 3-SAT problems being more complex to solve than 2-SAT problems. However, it turns out that not only are many interesting problems of the 3-SAT type, but these problems form a class of problems that can be reduced to one another. In other words, all of these problems share the same structure; developing a solution method for one of them, up to isomorphism, means having a solution for each of them.

Stephen Cook and Leonid Levin are credited with having simultaneously demonstrated in 1971 that any valid solution method for 3-SAT can be adapted to the methods of the NP class. This "transposition" itself is done in polynomial time, which means that the transposition is not more complex to execute than the solution method itself (which, until proven otherwise, is exponential). From a formal standpoint, this result forms the foundation of the concept of NP-completeness, of which the 3-SAT problem is the quintessential archetype. A problem is said to be NP-complete when it satisfies the following two properties:

- **It belongs to NP:** The problem P must be such that it is "easy" to verify whether a given answer s is a solution to this problem P. By "easy," I mean that the verification can be carried out in polynomial time. In other words, there exists an algorithm that, for any possible answer s, determines whether or not s is a solution to the problem; the time complexity of this algorithm is $\mathcal{O}(n^k)$, where k is a positive integer.

- **It is transposable to NP problems:** The problem P must be such that any problem P' that satisfies condition (1) can be reduced to problem P via a polynomial transformation. This means that if we have a solution s for problem P, there exists a function f that allows us to find the solution s' for problem P'. This function must be computable in polynomial time. In other words, it must have complexity $\mathcal{O}(n^k)$. Intuitively, we would say that the problem P is at least as hard as all problems in the NP class.

It is important to note that not all problems in the NP class are NP-

complete. Some problems in NP are not reducible to other problems in this class, which means that finding an efficient solution algorithm for these particular problems does not teach us anything about the other problems, like 3-SAT.

Conversely, there may be problems that are not in the NP class but whose solution method can be transposed to solve all problems in the NP class. A problem that satisfies this property is said to be NP-hard, rather than NP-complete. NP-hard problems must be such that finding a solution or verifying a solution are, in both cases, equally difficult operations. To date, no proof has been provided that NP-hard problems are not, ultimately, NP problems. Indeed, the fact that we do not know of an algorithm to quickly verify their solutions does not mean one does not exist[16].

If there is one piece of knowledge that has undoubtedly strengthened over time, it is this list of NP-complete problems. Since the 1970s, a number of problems, considered complex, were known in discrete mathematics, most often represented using graphs, sets, or even systems of linear equations: resource allocation, determination of an optimal route, filling a certain volume, etc. In 1972, Richard Karp showed that 21 of these problems are NP-complete and, thus, reducible to one another. Today, we know so many that it would be difficult to list them all. They span graph theory, game theory, equation solving, Euclidean geometry, and continuous optimization. I will give a few examples. Some are chosen from Karp's original list, while others are more recent:

- **The Traveling Salesman Problem:** A merchant wishes to visit a given set of cities. He knows perfectly the distance between each of these cities relative to one another. These distances are not calculated as the crow flies, but rather based on the actual length of the roads connecting them. He seeks to plot a route such that each city is visited at least once and the total number of kilometers traveled is less than a given integer k.

- **The Knapsack Problem:** A hiker wishes to take a certain number of items in his backpack, but the size of the backpack is limited. He assigns each item a value and a weight. He seeks to maximize the total value of the items he places in the backpack, without exceeding its weight capacity. There exists a multidimensional variant of this problem, where each item has a number d of parameters (weight, width, length, height...) and where the backpack itself has a different capac-

[16] In general, it is easier in mathematics to prove the existence of an object than its non-existence. This curse of non-existence is responsible for the fact that very few things can be stated with certainty regarding the algorithmic complexity of combinatorial problems.

ity for each of these parameters.

- **The Graph Colouring Problem:** A data analyst wishes to assign a colour or a label to each member of a social network such that two people who are friends on this network do not have the same colour or label. One can represent the social network by drawing a point for each registered person and connecting by straight lines the points corresponding to people who are friends with each other. In the case where none of the straight lines thus obtained intersects, four colours are sufficient to solve the problem. This result, known as the Four Colours Theorem, has a practical application, familiar to all schoolchildren: four colours are enough to colour the countries of a map such that no two neighboring countries share the same colour.

- **The Vertex Cover Problem:** King Minos wishes to post guards in his labyrinth in such a way that, with the minimum number of people, the entire network of corridors is covered. The guards must not move; they remain in fixed positions. However, they can look in all directions. The modern version of this problem involves placing cameras in a building to prevent James Bond from sneaking in unnoticed.

- **The Knight's Tour Problem:** A chess player places a knight on a random square of the game board. He wishes to visit all 64 squares without ever passing through the same square twice. Under the pen of Georges Perec, this problem can also be expressed as follows: how to visit all 100 locations of a 10 x 10 building without ever passing through the same location twice? The solution gave birth to the book *Life: A User's Manual*, where each chapter is dedicated to a room in this building[17]. But if you are more of an architecture enthusiast, you will likely prefer Euler's formulation: is it possible to wander in the city of Königsberg and return to the starting point, crossing each bridge exactly once? This problem is slightly different, as there is an additional constraint: returning to the same location.

- **The Minesweeper Game:** This famous single-player game on Windows, originally invented to help users familiarize themselves with the mouse, consists of a grid. Each square may either contain a bomb or a number corresponding to the number of bombs in the adjacent

[17] Georges Perec actually makes a small mistake between the 63rd and 64th move, but it is probably intentional since the correction is simple. See more on this at https://www.danthomasson.com/georges-perec.html.

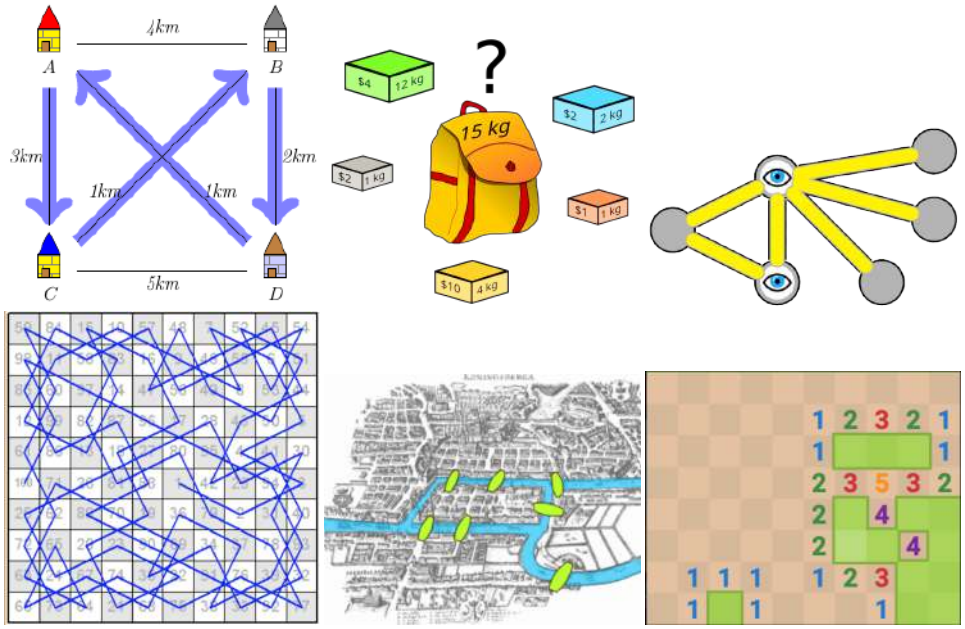


Figure 6.2: From left to right: (1) The Traveling Salesman Problem, (2) The Knapsack Problem, (3) The Vertex Cover Problem, (4) Georges Perec's 10x10 Knight's Tour Problem, (5) The Seven Bridges of Königsberg Problem, and (6) The Minesweeper Game.

squares. At the start of the game, the squares are hidden. The player can reveal them one by one: if he hits a bomb, he loses the game. Anyone who has played this "sport" realizes that sometimes you must "take a chance." But is it possible to win the game without relying on luck? Richard Kaye, from the University of Birmingham, demonstrated in 2000 that any given configuration of the game defines an NP-complete problem[18]. This means that if you want to analyze the game to avoid relying on "chance" as much as possible, you will need to solve an NP-complete problem. The proof of this intriguing result consists of showing that any 3-SAT problem can be constructed from Minesweeper[19].

[18] KAYE R., *The Minesweeper is NP-complete*, The Mathematical Intelligencer, Volume 22, number 2, 2000, pp. 9–1.5

[19] To be more precise, Richard Kaye demonstrates that Minesweeper can be used to construct logical gates such as "Negation," "And," "XOR" (exclusive or). By combining these gates, one can create a Minesweeper configuration corresponding to a logical proposition in conjunctive normal form. Any solution to Minesweeper is thus a solution to this 3-SAT problem.

It is easy to observe that these problems touch on many practical applications. Solving them with an Expert System amounts to solving 3-SAT. Currently, there is no algorithm capable of solving them in polynomial time, meaning that the computation grows exponentially with the size of the problem. This is one of the reasons, though not necessarily the most serious, why the project of reducing artificial intelligence to the solving of logical equations encountered serious difficulties starting in the 1970s. After an initial period of enthusiasm for Expert Systems, the question arose: which alternative methods could be used to address the problems in the NP class?

Two general strategies can be considered. The first, and the simplest to conceptualize, is to abandon exact solutions in favor of approximate solutions. In some cases, it is permissible to relax the constraints imposed on the solution in order to expand the search space. This will, of course, depend on the applications, but it seems that the human mind is capable of tolerating inaccuracy to navigate complex contexts, where machines refuse to proceed further. It may therefore be that tolerance for inaccuracy or ambiguity is a valuable resource for intelligence, understood in its broadest sense.

The second approach, entirely complementary, relies on an essential property of NP problems. It has been said that it is always quick (polynomial time) to verify an NP solution, but difficult (exponential time in current algorithms) to find it. Nothing prevents, therefore, an intelligent mind from trying to guess the right answer using statistics or experience. It will surely be wrong often, but by refining its approach over time, and provided the problem admits enough solutions, it is not impossible that it will eventually converge towards something robust.

This adaptive strategy, which does not conceive of the problem outside of its learning context, is also motivated by the fact that, regardless of the resolution method adopted for solving an NP problem, one must always start with an initial candidate and optimize it. In many cases, the choice of a good candidate will allow even a mediocre algorithm to converge in linear time, whereas a poor candidate will cause even the best algorithm in use to become entangled in calculations of exponential complexity.

Is it then not conceivable, by feeding the machine with enough known examples, to allow it to gradually improve its ability to "get lucky" and choose the "right" initial candidate? It seems that humans do not proceed differently when faced with a difficult task. Consider the labyrinth problem of Minos mentioned earlier. This problem involves placing a number of guards in the corridors to monitor the entire structure. If Minos asked his advisers how to proceed, and if they were inexperienced strategists, they might come up with answers that are just as poor as if they proceeded randomly. But if experience were on their side, it is plausible to think that the advisers would not submit overly poor proposals to their sovereign. A careful study would then determine which of them is the best—and refine it as much as possible—but, so to speak, we would start with good raw

material.

The learning ability I have in mind, based on experience, may be akin to Socrates' Daimôn? The Daimôn indeed whispers to the philosopher's ear about impractical paths. It does not point out what is good to do, but it awakens each time the foot is set on a path with questionable moral consequences. Similarly, experience or intuition is a small voice within us that guides our gaze. It does not tell us how to act in the face of a problem, but it leaves us with an uncomfortable feeling, an indefinable something, when our mind turns to dubious approaches. The more a person acquires practice, the more he refines this voice within. He will no longer submit to his judgement anything that does not seem worthy of interest. And even if the most experienced strategist sometimes misses the ideal solution, at least he will not waste his time studying the wrong ones.

All these considerations have led artificial intelligence down a new path. Instead of describing the course of action to algorithms purely from a Top-Down approach, it has become increasingly interesting to provide them with the nourishment to grow by themselves. This bottom-up approach, which starts from raw data and gradually ascends towards rational computations, is at the heart of neural networks. The model of such a machine is no longer the impartial judge, who decides based on logical laws, but the self-adaptive living organism; the one that accumulates knowledge dynamically. This new field of research is called *Learning Intelligence*. This is what will be discussed in the next chapter.

6.7 Epistemological Critique: Between Platonism and Logicism

My ambition in this section was to demonstrate how top-down strategies, most perfectly embodied by expert systems, rely on analytical judgement without managing to capture the *a priori* forms of human knowledge. When Herbert Simon declares that such systems will, within twenty years, be able to perform any cognitive task within the capacity of a human being[20], he implicitly assumes that human intelligence can be reduced to a set of logical deductions, or that the hypothetico-deductive method could at least replicate, if not surpass, human abilities.

It is, however, interesting to note that the analytical judgement referred to in Herbert Simon's General Problem Solver is not the same as the one identified by Plato. On the contrary, it is closer to calculating reason. It is a speculative machine detached from any semantics. Its thought operations reduce to writing symbols on paper and deriving rigorous conclusions from them.

[20] He writes around 1960: "Machines will be capable, within 20 years, of doing any work that a man can do." (The New Science of Management Decision)

Now, Plato identified two rational faculties: *dianoia* and *noésis*. The former consisted of making logical inferences from certain hypotheses, somewhat in the manner of Euclid's Elements of Geometry. *Noésis*, on the other hand, was the contemplation of the anhypothetical, or, in other words, that which lies beyond all hypotheses. This anhypothetical is often assimilated to the idea of the Beautiful, the Good, and the True. However, it may be that this is merely an analogy in Plato's thought[21].

To contemplate a pure form, for Plato, is to contemplate them all. In the intelligible realm (*topos noetos*), the Ideas are so intricately interwoven that they form a coherent system. This realm is the stupor of the geometer who, all at once, is struck by the beauty of mathematical figures, the harmony of their relations, symmetry, concision, and the ineffable elegance of the laws that govern them. Plato's philosophy follows an ascending and descending dialectic. First ascending, because it starts from sensible forms and seeks to understand the Ideas they participate in—by observing drawings of triangles in the sand, one aims to grasp the Idea of the triangle itself. It is also descending, because it seeks, starting from the most general Idea, to establish distinctions of species between concepts.

Expert systems seem to have taken this dual movement seriously, but by bypassing this *noésis*. They do indeed have foundational principles: they are the axioms of the theory and logical laws. It is even permissible to seek new truths based on these principles, either in an ascending or descending manner. However, the use of the words "ascending" and "descending" is precisely the opposite of what Plato saw in them. The ascending approach of expert systems indeed begins with the first principles to deduce true propositions, creating distinctions. The descending approach, on the other hand, starts from propositions and proves them by reducing them to axioms. In some cases, it also allows for the addition of clauses to the theory.

Thus, it seems as if the expert system must already possess the first principles in order to "think." Instead of observing the sensible world to derive intelligible truths, it operates in a world disconnected from experience. The machine can indeed receive data to draw conclusions, but these data must be logically formatted. It is not possible to ask the machine to analyze a photo and create its own categories, as is done today with neural networks. Furthermore, there is no capacity for self-learning: unless you modify the rules, the system remains, eternally, the same.

In compensation, the programmer has full control over the machine's behavior. Its conclusions are traceable and justifiable: it is possible to predict

[21] There is a debate opposing two interpretations. According to the ontological interpretation, the Idea of the Good is truly the anhypothetical. According to the epistemological interpretation, Plato compares the anhypothetical to the Idea of the Good only by analogy, just as he compares the Idea of the Good to the Sun. See Roux Sylvain, *Chapter III. The Good as Principle in the Republic*, in *The Search for the Principle in Plato, Aristotle, and Plotinus*, Vrin, 2004, pp. 69-116, https://books.openedition.org/vrin/5432?lang=fr

how it will behave in many contexts and understand which principle, which clause, was responsible for its actions. However, there is no contemplation of the truth, in the sense of an exercise aimed at transforming one's gaze upon the world. The machine does not seem to educate itself in a moral or intellectual sense, to gradually ascend in understanding an ideal, an intelligible truth.

In summary, one could say that the forgetting of the *Entendement*, which began with the crisis of mathematics, has been accentuated in the early programs of artificial intelligence, turning Reason, no longer into the faculty of judgement regarding the great Ideas, but into an aptitude for calculation. This reduced version of Platonic reason is what I will call numerical reason, or even: modern rationality. What reason loses in terms of epistemology, it gains in normativity. What do I mean by this? That the programmer of the expert system has complete control over what is true and false. He encodes the laws of thought and the axioms of the theory; moreover, he checks the answers. Expert systems are traceable (one can verify the steps of reasoning) and verifiable (one can determine if the answers are true or false).

Despite these obvious advantages, one must ask how programming paradigms have evolved and whether they have succeeded in integrating the sensible world into their calculations. As Kant indeed said, reason should not spin in vain. It needs to rely on objects of space and time. It must start from experience and ascend to concepts in a bottom-up fashion.

What to take away from this chapter

The first programming strategy to emerge is referred to as the *top-down* strategy. It consists of determining rules of reasoning, encoding them into an algorithm, and executing them with a machine. This approach can be seen as a procedure for solving logical equations, written in the form of clauses such as "A and B", "A or B", etc.

This procedure is called an expert system, but its expertise is limited to the rules that have been taught to it. The expert system, in fact, has several shortcomings:

1. It lacks autonomy or self-criticism.

2. It is quite unsatisfactory when faced with certain resource-intensive problems, particularly NP-complete problems.

3. It reduces rationality to the mere manipulation of symbols. It overlooks the aesthetic and universal dimensions of human reason, conveying a vision of rationality based on calculation.

In the next chapter, I will ask whether there exists another programming strategy.

Chapter 7

Bottom-Up Approach: Neural Networks

Emotions are intertwined in the neural networks of reason.

Antonio Damasio

For a long time, expert systems represented the archetype of artificial intelligence. These software systems are the Sherlock Holmes of human knowledge: from a small number of fundamental principles, they are able to deduce, infer, and exclude possibilities that might not necessarily have been considered by human minds. The rules of reasoning and principles used by these machines must, however, be encoded by a programmer. These rules must be expressible in a precise and operational manner. They must also have consensus from qualified individuals in the field.

However, as early as 1943, long before expert systems existed, McCulloch and Pitts proposed an alternative approach to artificial intelligence. In a paper titled "A logical calculus of the ideas immanent in nervous activity", they hypothesized that human learning abilities are linked to a specific way in which neurons receive, process, and transmit information. From there, they sought to understand and model the mathematical properties of the neuron based on known data about the physiological functioning of the brain[1].

> *Due to the "all-or-nothing" nature of neuron activity, neuronal events and their relationships can be treated using propositional logic. It turns out that the behavior of each network can be described in these terms, with the addition of more sophisticated logical processes in the presence of networks containing cycles.*

[1] McCulloch, W.S., Pitts, W., *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics 5, 115–133 (1943).

Despite a considerable simplification of the biological processes involved in the functioning of the brain, and perhaps because of it, McCulloch and Pitts' model remains a crucial starting point in the development of bottom-up algorithms. Numerous variants have since been introduced by authors such as Hopfield (Nobel Prize 2024), Kohonen, or Hérault-Jutten. However, it is now widely accepted that mathematical models of neural networks, while convenient for generating machine learning, do not represent a valid theory of neurological processes.

In this chapter, I will first describe the functioning of the biological neuron, after which I will introduce McCulloch and Pitts' mathematical model. I will then discuss various improvements to the initial model, specialized in image perception or information retention. I will conclude by illustrating the fundamental differences between neural networks and expert systems using the software AlphaGo. This will lead to some epistemological conclusions in the next chapter.

7.1 The Biological vs Formal Neuron

The dominant histology of the 19th century represented the brain as a mass of fused cells. Since the work of Santiago Ramon y Cajal, Nobel Laureate in 1906, it has been accepted that the neuron is an autonomous functional unit. I will limit myself here to providing some general characteristics of its functioning. I will mainly focus on the aspects of the neuron most relevant to understanding the links between neurophysiology and theoretical computer science. I will therefore leave aside terminology concerning different types of cells (Purkinje cells, astrocytes, ependymocytes...) or the exact relationship the nervous system has with the rest of the organism (nociception mechanism, role of the spinal cord, or the famous stomach neurons...).

The analogy often used to describe how neurons function to one's grandmother is to imagine a mass of tangled wires (it should be acknowledged, for the sake of this metaphor, that your grandmother is not electro-phobic). Each wire has a receptive end and an emitting end. The receptive ends are connected to the emitting ends. However, there is a gap between the wires, so the current cannot flow until contact is established. This can happen when a drop of water fills the gap between them. These drops are secreted by the ends of the wires themselves in the presence of an electrical potential. Some of these drops have shapes that allow them to establish contact more efficiently between certain wires rather than others. When two wires make contact via a drop of water, an electric discharge is likely to occur. Thus, an electrical signal is transmitted from one wire to another.

In this metaphor, the wires represent individual *neurons*, the space between the wires represents the *synaptic gaps*, and the drops of water represent the *neurotransmitters*. Neurons are nerve cells composed of a nucleus,

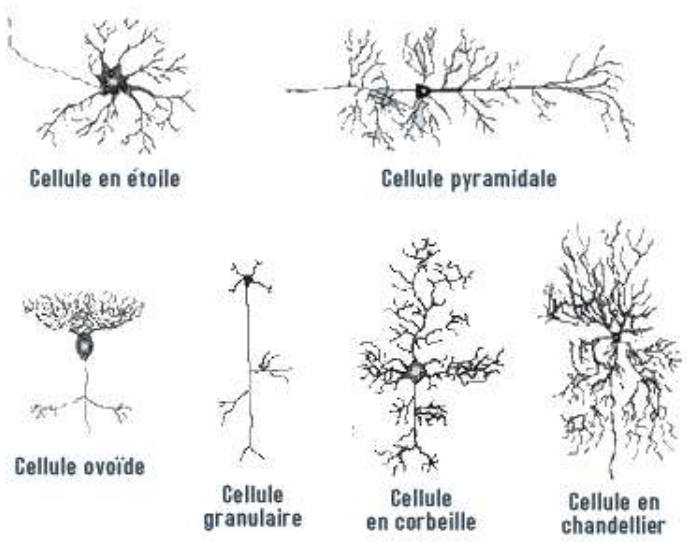


Figure 7.1: Neurons can take different shapes.

cytoplasm, and membrane. The nucleus contains genetic material, the cytoplasm is rich in proteins, and the membrane ensures the coherence of the cell. The neuron, and this is its specificity, also has tentacular extensions, giving it the shape of a star, a pyramid, a basket, or a chandelier. These extensions are of two types. Those that receive electrochemical signals from other cells are called dendrites. The extensions that spread the signal from the neuron to other cells are called axons. There is thus a receptive end and an emitting end (see figure 7.2).

Neurotransmitters, represented by a drop of water in our analogy, are molecules synthesized by neurons as well as by other cells in the nervous system, such as glial cells. There are over a hundred known neurotransmitters, classified into different families: amino acids (glycine, glutamate...), monoamines (dopamine, adrenaline, serotonin...), and peptides (opioids...), to name a few. These neurotransmitters are often associated with specific physiological and psychological roles. Dopamine, for example, is described as responsible for pleasure; adrenaline contributes to stress management; serotonin is the happiness hormone, etc.

However, it is important to be cautious of these simplifications—what some call neuromyths. While it is true that depression treatments, for example, involve chemically increasing serotonin levels, it is scientifically dangerous to reduce each neurotransmitter to a unique role. When human concepts (happiness, pleasure, energy...) are associated with physiological units (serotonin, dopamine, adrenaline...), it is assumed that there is a direct equivalence between our psychological variables and observable cells. Yet, there is no reason to believe that the map we use, as humans, to navigate our

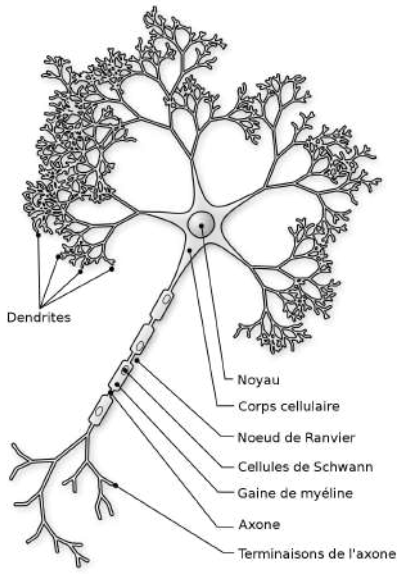


Figure 7.2: Biological diagram of the neuron.

psychosocial existence is similar to the map nature uses to assign functions to the different cells of our organism.

As an illustration, a deficiency in adrenaline in the substantia nigra, a nucleus located in the brainstem, can cause Parkinsonian symptoms (hand tremors); while an excess of this same adrenaline in the mesolimbic pathway can lead to schizophrenic symptoms (psychotic episodes). This dual effect of adrenaline is far removed from the caricatured representation commonly attributed to it, when people say they "love adrenaline" or talk about an "adrenaline rush." This phenomenon is well-known to pharmacologists: when prescribing psychoactive drugs, they are careful to avoid falling from Charybdis to Scylla, causing schizophrenic reactions in a patient suffering from Parkinson's disease.

With these reservations in mind, what can we say about the role of neurotransmitters in neuronal activation? Neurotransmitters generally act as catalysts. The more they are present in a synaptic gap, the greater the chances of allowing the current to pass from one neuron to another. When a neuron A is excited, it releases neurotransmitters into the synaptic gaps of neurons $B_1, ..., B_n$ to which it is connected by dendritic extensions. This release of neurotransmitters is generally insufficient, on its own, to excite these neurons $B_1, ..., B_n$. However, when several neurons incident to neuron B are excited simultaneously, B will become excited in turn. In this way, neuronal excitation will propagate throughout the brain.

In other words, and this is a fundamental law, a neuron will only activate if it is in contact with enough other excited neurons. This rule, very

important for what follows, means that variable electrical pathways can be created from one neuron to another. The learning of the brain consists of associating certain neuro-electrical paths with specific contexts, so that every time a similar situation occurs, a similar path is created. This phenomenon explains why we memorize information more easily when it is linked to information already known, rather than entirely new information. This idea of learning through statistical associations of neuron activations is, in fact, the foundation of the mathematical concept of the "formal neuron".

7.1.1 The Formal Neuron

McCulloch and Pitts' neuron is a simplified model of the biological neuron's function, which omits the complexity of physiological processes to retain only four key ideas:

1. The neuron is a functional unit that receives a signal from one or more neurons (called input neurons) and sends a signal to one or more neurons (called output neurons);

2. The signal sent is represented as a binary number or a real number;

3. When the sum of the signals received by a neuron, the potential p, exceeds a certain threshold θ, this neuron becomes active and sends a signal to the output neurons it is connected to;

4. The signals emitted by the input neurons are multiplied by a synaptic weight, which models the importance of the connexion between the two neurons.

From a mathematical perspective, the neuron receives inputs $x_1, x_2, ..., x_n$ and calculates the potential p. This potential is given by the weighted average of the inputs: $p = x_1W_1 + x_2W_2 + ... + x_nW_n$, where $W_1, W_2, ..., W_n$ are the "synaptic weights" specific to the neuron. These parameters correspond, in the biological model, to the receptivity of the dendrites. The higher the weight W_i, the more important the associated signal x_i is in the neuron's response. If, on the other hand, W_i is zero, the signal x_i is not considered: it is as though there were no connexion between the neuron and the input in question.

The response of a formal neuron is binary: either the neuron activates, or it does not. Mathematically, this means that the output signal from the neuron, denoted y, is a function defined for all real values, with outputs of -1 (inactivity) or 1 (excitation). The value -1 is observed if the action potential p is below a threshold θ. The value 1 occurs when the action potential exceeds or equals the threshold θ:

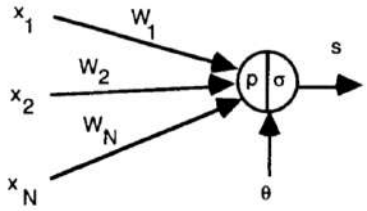


Figure 7.3: The formal neuron of McCulloch and Pitts (1943).

$$y = \begin{cases} 0 & \text{si } p < \theta \\ 1 & \text{si } p \geq \theta. \end{cases}$$

It is worth noting that some formal neurons do not return the response -1 or 1, but rather 0 or 1. This choice depends on the applications at hand and certain conventions regarding the variables used. It is also possible to use a decision function to produce real-valued responses between -1 and 1. This approach is particularly interesting when the expected response is not an all-or-nothing type. It is also a good method if neuron A produces a signal that will be read by neuron B. The outgoing signal from A can be more or less significant, which will activate neuron B to varying degrees.

The McCulloch and Pitts model can therefore take different forms depending on the number of inputs $x_1, x_2, ..., x_n$, the values of the weights $W_1, W_2, ..., W_n$, and the decision function linking the weighted sum of inputs to the output y. The number of inputs primarily depends on the problems being addressed: in image recognition, there will be three inputs per pixel, one for each primary colour; in data classification, there will be one input per data point, and so on.

The decision function, for its part, will be constructed with regard to the problem being solved. If the output of the neuron is to provide a binary value, such as for the opening or closing of a door, a Heaviside function must be used, which produces only the response -1 or $+1$ (see Figure 7.4); but if the output is a real value, such as when determining the ambient temperature using sensors, one might opt for a sigmoid function, whose input/output characteristic is, by the way, closer to that of the real neuron (see **??**).

In a formal neural network model, learning consists of calculating the values of the synaptic weights $W_1, W_2, ..., W_n$ based on available examples. In contrast to top-down methods, it is not necessary to heuristically isolate the fundamental principles for solving the problem. The process is blind, providing the machine with the formal parameters for learning. These parameters are constructed in such a way that the model's responses coincide with the known solutions to the problem. The more precise and numerous these data are, forming a statistically varied sample, the more efficient the

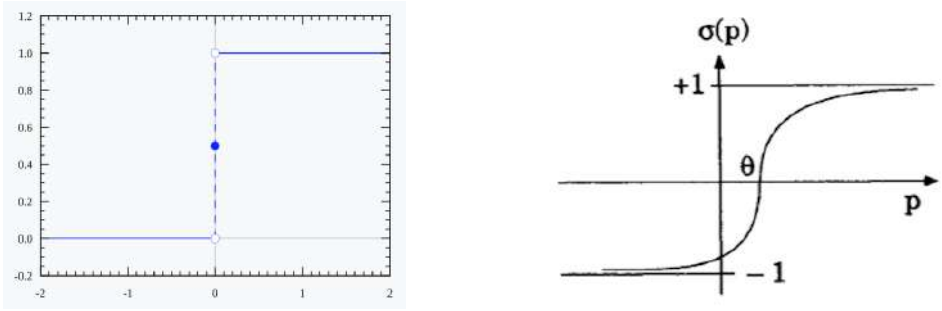


Figure 7.4: The Heaviside function, on the left, can only take values -1 or 1. The sigmoid function, on the other hand, produces continuous outputs between these two extremes. On the graph to the right, the decision function is $2sig(x) - 1$, so that the values are between -1 and $+1$.

resulting model will be. It is therefore not uncommon for the machine resulting from this approach to adopt behaviors that are sometimes difficult for a human to decipher. This is referred to as intrinsic opacity.

As an illustration, imagine a formal neuron tasked with deciding whether Buridan's donkey should eat hay or oats. This neuron takes two inputs, x_1 and x_2, quantifying its appetite and its cold symptoms, respectively. The higher x_1 is, the hungrier the donkey is. The higher x_2 is, the more the donkey is sensitive to coughing fits. These two values x_1 and x_2 have no upper limit, but must be equal to or greater than zero.

The donkey's response — to its great dismay — is binary: either $y = 1$, in which case the donkey eats hay; or $y = -1$, in which case it eats oats. The decision function can thus be taken as:

$$y = \begin{cases} -1 & \text{if } p = W_1 x_1 + W_2 x_2 < 0 \\ 1 & \text{otherwise.} \end{cases}$$

This example corresponds to the perceptron imagined by Frank Rosenblatt in 1957, which aims to separate a set of data into two disjoint classes. In this case, these classes correspond to "eat hay" or "eat oats". The synaptic weights W_1 and W_2 must be chosen in such a way as to predict the donkey's response. One might assume that hay is more nourishing than oats, so that the hungrier the donkey is, the more it will turn to hay. Similarly, one could estimate that oats cause fewer sneezes than hay, so that the more the donkey is sick, the more it will turn to oats. A simple solution would be to set W_1 positive and W_2 negative. However, there are many other possible choices. Which are the most appropriate? According to what criteria?

7.2 Hebb's Rule: Birds of a Feather Flock Together!

I have given the definition of the formal neuron by McCulloch and Pitts and compared it to the biological neuron. I have shown that the whole challenge of artificial neurons lies in creating relevant associations between the inputs $x_1, \ldots, x_n$ and the output y. The question that arises at this stage is to understand how exactly the synaptic weights W_1 and W_2 must be set to enable the neuron to perform certain learning tasks? In the example of Buridan's donkey, we wonder how the formal neuron of our donkey can allow it to make better choices. To achieve this, we would like to use the data already available to us: namely, all the past experiences in which the donkey, after feeling a certain hunger and a certain degree of cold, turned to one or the other edible substance. We would like to repeat the choices that turned out to be wise and inhibit the others.

In his book *The Organization of Behavior*, published in 1949, Dr. Hebb formulates a rule about the behavior of neurons that has had a significant impact on solving the problem that concerns us. This rule, now known as the "Hebb's rule," states that when two neurons excite together, they remain linked to each other:

> *When the axon of cell A is sufficiently close to excite cell B and it repeatedly or permanently participates in its activation, a process of growth or metabolic change takes place in one of these cells, or both, so that the efficiency of A, as an excitatory cell of B, is increased.*

This physiological law, when transposed to the McCulloch and Pitts model, suggests increasing the value of synaptic coefficients only between formal neurons that have synchronized activity, and only those. To understand how this can be done in detail, I would like to illustrate the method with a single formal neuron and then extend it to a complete network.

7.2.1 Isolated Neuron

Let us begin by considering that our Buridan's donkey, just out of the hands of God, is endowed with a single, unique neuron. This neuron is connected to two inputs: hunger x_1 and cold x_2. It produces a food-related response y that equals 1 (eating hay) or -1 (eating oats). In accordance with Hebb's rule, the neuron will attempt to reinforce existing habits. If our animal often eats oats when it is cold, the neuron must therefore favor the response $y = -1$ when x_2 is high.

From a mathematical point of view, we can represent all the past experiences of our donkey, its previous preferences, using three points $x_1, x_2,$

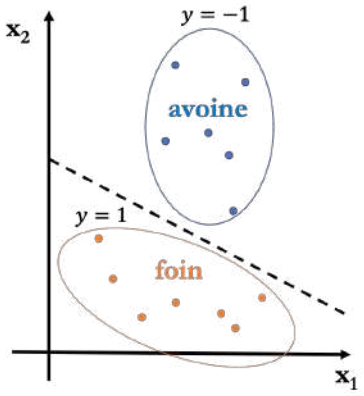


Figure 7.5

and y. The inputs x_1 and x_2 can be placed in a two-dimensional plane (see figure 7.5). These points (x_1, x_2) are coloured red when associated with the response $y = 1$ and blue when associated with the response $y = -1$. Under this representation, the role of the neuron is to draw a boundary between the red and blue points using a statistical regression line, indicated by a dashed line.

After drawing the boundary, we say that the neuron has completed its learning phase; it is ready to predict the donkey's future behaviors. When it receives an unknown input (x_1, x_2), it will respond 1 if it is located above the line and -1 if it is below the line. In other words, the neuron uses past experiences to predict future behaviors.

To obtain this famous regression line, which serves as the boundary between $y = 1$ and $y = -1$, several methods exist. The most well-known, which will no doubt remind you of your statistics classes, consists of finding the line D that minimizes the sum of the squared distances. Intuitively, D is the line that passes exactly in the middle between the set of blue points and the set of red points. The points used for the neuron's learning are said to belong to its learning base. New points, on the other hand, are predictions.

The statistical regression line is generally an effective separator, but to calculate its slope, it is necessary to have the entire set of red and blue points from the learning base. This means that every time a new data point is added to this base, the slope of the line must be recalculated in its entirety. However, the computation time grows with the number of points. In the presence of big data, sometimes containing millions of points, our neuron will be ill-equipped to absorb new information. It is therefore not flexible enough to adapt to new contexts, as one would want in the bottom-up approach.

To address this difficulty, the ADALINE model, created in 1960 by Bernard Widrow and Ted Hoff, takes a completely different approach. Rather than looking at the entire set of points in the learning base all at once, it

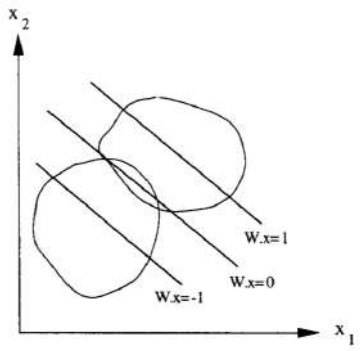


Figure 7.6: Example of non-linearly separable classes.

considers them one by one. It uses the first points to trace a very rough separation line, necessarily erroneous. Each time a new point is considered, two scenarios are possible. Either this point lies on the correct side of the line, in which case nothing happens. Or the point is on the wrong side of the line, in which case the neuron slightly changes the orientation of this line to minimize the error. Gradually, the line will converge towards the least squares method. Therefore, ADALINE is as effective as a standard statistical regression, but with dynamic learning.

If the learning is good and the data has a good structure, the neuron is able to retrieve the responses $y(i)$ associated with the points $x(i)$ from its learning base. There is, therefore, an automatic association of the learned inputs with the corresponding outputs. What interests us, however, is what the neuron will do when the input is not exactly equal to $x(i)$, but is close? Indeed, most real-life situations are not exact copies of the learning repertoire, or we could not talk about the intelligent adaptation of the neuron, just as we wouldn't say a chess player is intelligent if they only play memorized opening moves. In other words, the entire interest of bottom-up artificial intelligence lies in the fact that the neuron will "improvise" in the face of new situations, attempting to relate the unknowns as closely as possible to what it already knows.

Since the association equations we have established are linear, the neuron's responses will be linear combinations of the memorized data. Technically, we say that the neuron performs first-degree statistical regression. Although this represents a good approximation, this approach can leave much to be desired in several aspects. First, there are data classes that are not linearly separable, meaning no linear method can determine without error whether an input x should be associated with output s_1 or output s_2 (see graph). To put it in more figurative terms—and to use the terms of the previous example—this means that there are situations in which Buridan's donkey cannot decide, without risk of error, whether hay is better than oats

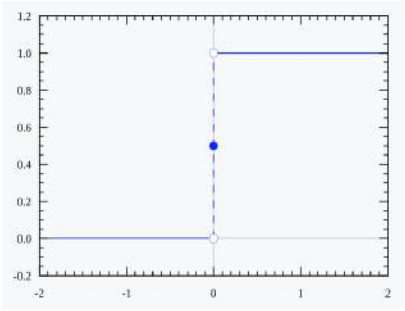
Figure 7.7: If the regression is not robust enough, it may fail to properly identify the Eiffel Tower when there is a change in lighting.

(see figure 7.6).

Another problem with linear methods is that they are less robust to changes in context. Ideally, we would like the formal neuron to be resilient: when exposed to pairs of points (x_1, x_2) it has not encountered before, we want its prediction to remain robust. A lack of resilience is evident, for instance, when a small change in light parameters invalidates an image recognition algorithm, causing it to fail to recognize the Eiffel Tower in the dark (see figure 7.7).

In technical terms, if the vector x passed as input is close to a memorized vector $x(i)$, the neuron should produce an output y identical or close to the memorized output $y(i)$. In the case of linear statistical regression, unfortunately, the risk of error increases as we move away from known situations. To improve the robustness of predictions, non-linear methods should be considered. These can be obtained by using a non-linear decision function.

In the perceptron we have considered so far, let us recall, the responses are either $y = 1$ or $y = -1$. This choice follows a "staircase" function, also called a Heaviside function, depending on whether the sum $x_1W_1 + x_2W_2$ is less than or greater than a certain threshold θ. However, it is possible to take, for the decision function, a sigmoid (figure 7.8). In this case, the values of y will vary between -1 and 1. It is also allowed to use the response produced by a first neuron as an input signal for a second neuron. By juxtaposing neurons and using non-linear decision functions, we can refine the robustness of statistical regressions. These multi-layer neural networks



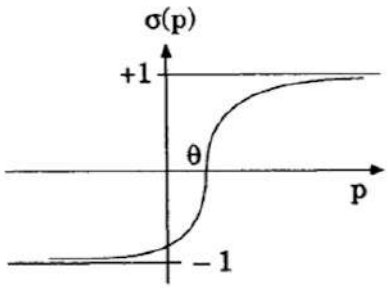


Figure 7.8: The Heaviside function, on the left, can only take the values -1 or 1. The sigmoid function, on the other hand, produces continuous outputs between these two extreme points.

will be the subject of the next sections[2].

Before delving into the subject, I would like to point out that the most elegant methods are not always those that calculate exact parameters, but those that adapt organically. Computationalist approaches, which equate the human mind with the brain and the brain with a computer, assuming that neuronal decisions are the result of analytical calculations similar to those performed by an electronic circuit, are indeed regularly challenged by neuroscientific advances. This is eloquently expressed by Francisco Varela in an interview with Herve Kempf[3] :

> *The notion of a "neural" computer is not self-evident, because a computer, strictly speaking, is a symbolic system. The current movement known as cognitivism, or computationalism, operates with the assumption that the cognitive level is autonomous. Their hypothesis is that there is a difference between the levels of implementation, between the software and the hardware. So, talking about the "neural" computer is already diving into the confusion of half the researchers in the field, for whom what the brain does is not directly relevant to understanding cognition. If you talk to a computer scientist, they do not care about the specific computer they are working on; what interests them is the structure of their program. Similarly, those who focus on computation itself are not*

[2] I only provide a brief characterization of neural network techniques here. The curious reader interested in more detailed approaches used before the 2000s is invited to consult the excellent popular book by one of my mathematics professors at UCLouvain (Michel Verleysen, François Blayot, *Les réseaux de neurones artificiels (Que sais-je ?)*, Presses Universitaires de France, 1996). For a more contemporary exposition, refer also to Stuart Russell, Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education.

[3] Francisco Varela, *La Recherche*, No. 308 April 1998, p.109-112. English translation is mine.

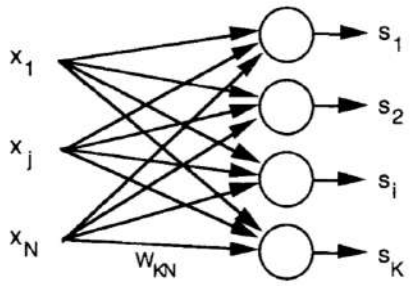


Figure 7.9: Single-layer neural network.

concerned with its physical substrate. They do not deny the importance of biology, they simply state that it is not their domain. Moreover, there are numerous ways to bridge the gap between computationalism and the brain. The question becomes how to analyze what a computation based on a non-symbolic machine is, one that is more comparable to a brain than to a computer.

7.2.2 Single-Layer Network

So far, we have considered an isolated formal neuron, whose output consists of a single decision such as "eat hay" or "eat oats". However, there are many applications where the output cannot be described by a single variable. One might think of an artificial intelligence tasked with drawing an image with 1024×1024 pixels, or an automatic flight control program managing the orientation of the aileron, the joystick, and the rudder. For such situations, the methods discussed previously remain unchanged, but for each new desired output, a new formal neuron is introduced. This neuron takes all the variables $x_1, x_2, ..., x_n$ that constitute the problem as inputs. It also has a specific synaptic coefficient for each of these inputs.

The rules for adjusting the coefficients W can follow the same approaches as those considered earlier. With k neurons, it is enough to perform the calculation k times for the k different outputs. Depending on the application's requirements, algebraic or iterative methods can be used. However, the processing power of a single layer of neurons is not different from that of a single neuron. It is simply a linear separator. The only difference is that the neural layer is a set of linear separators. To solve problems that are not linearly separable, the network's structure must be strengthened, notably to allow for the composition of layers.

7.2.3 Multi-Layer Networks

It is possible to add several intermediate layers between the inputs $x_1, x_2, ..., x_n$ and the outputs $s_1, s_2, ..., s_k$. These layers $C_1, C_2, ..., C_q$ are populated with

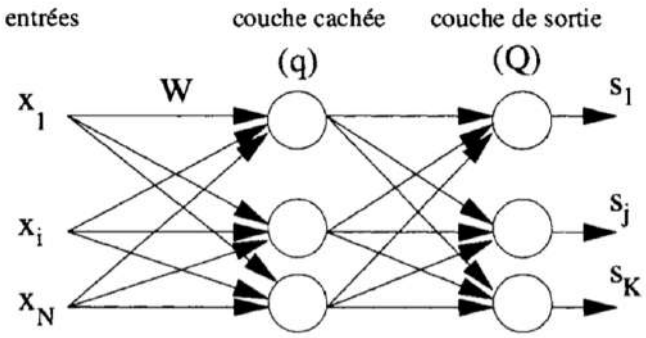


Figure 7.10: Multi-layer neural network.

neurons that take as input the outputs from the previous layer. Each of these neurons then produces an output that serves as input to the neurons in the next layer, and so on. Note that there is no requirement that these layers each have the same number of neurons. Nor must these neurons have the same decision function σ. To achieve a good generalization of the Adaline model, however, it is advisable to associate all neurons with a non-linear, differentiable function σ, such as the sigmoid or, very similarly, the hyperbolic tangent[4].

Note that the neurons in the intermediate layers do not necessarily have the same number of neurons as the initial or final layers. The number of layers, as well as the size of these layers, is a free parameter. Typically, these are set by an experienced programmer who, through trial and error, determines the best structure. Later, we will see that it is also possible to entrust this task to an algorithm, known as a self-learning algorithm, or more commonly, unsupervised learning.

Synaptic coefficients are introduced for all existing connexions in the network. Thus, there is one associated with each edge in the figure above. The problem that arises is, once again, how to choose the value of these parameters? This is not an easy question. In the networks we have considered so far, the approach essentially involved reducing, using a gradient descent method or least squares, the error observed between the output produced by the neural network and the desired output. However, in the case of a multi-layer neural network, it seems impossible to determine the error made by the neurons in the intermediate layers. The available data only specifies

[4]The hyperbolic tangent *tanh*, defined as the ratio of the hyperbolic cosine and sine, has a very similar form to the sigmoid function. For this reason, it is often used in neural networks. The sigmoid $sig(x)$ is given by the equation $\frac{1}{(1-e^{-x})}$, while $tanh(x) = \frac{cosh(x)}{sinh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^x(1-e^{-2x})}{e^x(1+e^{-2x})} = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{2-(1+e^{-2x})}{1+e^{-2x}} = \frac{2}{1+e^{-2x}} - 1 = 2sig(2x) + 1$. The hyperbolic tangent is thus twice as steep as the sigmoid function and is also shifted by one unit downward.

what the outputs should be for a given set of inputs $x_1, x_2, ..., x_n$ at the end of the process, at the level of the final neuron layer. The operations performed in the preceding layers should only aim to improve the network's final output.

To address this challenging issue, P. Werbos, D. Rumelhart, and Le Cun introduced, and successively refined, the method known as *backpropagation of the gradient*[5]. This method is based on a rather intuitive approach. Since only the output of the final layer can be used to define an error margin, it is necessary to assign "objective functions" to the intermediate layers that must be minimized. To do so, let us assume, without loss of generality and without worrying about notation, that the neural network has q layers of neurons denoted $C_1, C_2, ..., C_q$. We begin by associating with the penultimate layer of the network (layer C_{q-1}) an objective function f_{q-1} that depends only on the synaptic weights linking C_{q-1} and C_q. The function f_{q-1} should be such that, when f_{q-1} is minimized, the error observed in the outputs of the final layer is also minimized. In this way, we have a local objective for layer C_{q-1} that coincides with the global objective regarding the outputs of the network. To give the other intermediate layers a similar objective, we simply repeat the procedure: we associate a function f_{q-2} with layer C_{q-2}, whose minimum coincides with the minimum of f_{q-1}, and so on.

The main disadvantage of the backpropagation method is its very high computation time, which leads to a long learning process. For some applications, where the network can be carefully optimized before use, this is not of critical importance. But if the goal is to model human cognition, which continuously adapts to encountered problems, one may fear a lack of fluidity. To bypass this bottleneck, it is possible to use other minimization methods, such as the conjugate gradient method or approaches inspired by Newton-Raphson[6].

7.3 Deep Learning: Myth or Reality?

Rather than leaving the choice of intrinsic parameters of the neural network, such as the number of layers, the number of neurons per layer, the type of function to minimize, or the method of minimization used, to an experienced programmer, this choice can be entrusted to an algorithm. This path does not exempt the use of human judgement at the outset of the process, but it grants the machine greater autonomy. The machine is even likely to evolve iteratively into a form unplanned by its designer: its network may expand or contract organically, its synaptic weights can be adjusted as new data is

[5] D. Rumelhart, D. McClelland, *Parallel Distributed Processing. Exploration in the microstructure of cognition, vol I.* Cambridge, MIT Press.

[6] These optimization methods generally involve calculating the slope of the function and seeking where it vanishes. They were quite popular at the beginning of artificial intelligence but are no longer widely used.

introduced, and its structure, to put it briefly, is capable of evolving like a simulacrum of a living being.

One of the first evolutionary neural networks discussed in the literature is the one by Reilly and Cooper[7], named the Restricted Coulomb Energy (RCE). Unlike the algorithms considered so far, which aim to separate data categories either linearly or non-linearly, the RCE's role is to determine regions of influence. What I mean by this is that the algorithm does not aim to discriminate all the situations where the optimal response is 1 (eating hay) or -1 (eating oats), but rather to draw a kind of confidence zone around the cases where the correct response is known.

To provide a geometric illustration, one could say that the usual task of a neural network is to partition areas, or categories, within the space of possible data. The data space has as many axes as there are inputs $x_1, x_2, ...$ In the case of Buridan's donkey, for instance, it is a two-dimensional space with the degree of hunger on the abscissa and the degree of cold on the ordinate, but in more complex cases, the space may have hundreds of dimensions. By calculating the weights $W_1, W_2, ...$, the neural network determines boundaries within this space. In the case of a two-dimensional plane, these boundaries are straight lines, but in the more general case, it is preferable to speak of hyperplanes. If well-constructed, these boundaries must be such that when one is on one side, the output is $+1$, while on the other side of the boundary, the output is -1. If the network admits more than $+1$ and -1 as responses, there may be more complex and numerous boundaries. However, there is no region of the space that is not associated with a response from the neural network.

In the case of the RCE algorithm, however, we do not seek to partition the entire space of possible data into distinct categories; instead, we only start from the known points (the training data) and draw, around each of them, regions where it is likely that the correct response will be the same as that associated with the starting point. If Buridan's donkey, to revisit our canonical example, ate oats when $x_1 = 0.8$ and $x_2 = 0.2$, the algorithm might decide that, within a small radius around the point $(0.8, 0.2)$, the donkey should eat oats. This approach is called constructive, because the regions of influence are not obtained by partitioning the initial plane with hyperplanes, but by constructing increasingly larger zones around the known points.

From a procedural perspective, the algorithm begins by drawing an arbitrary radius (set by the RCE parameters) around the data it holds. If two zones collide, there are two possibilities. Either these two zones are associated with the same response ($s = 1$ ="eating hay"), or they are associated with distinct responses ($s = 1$ ="eating hay" and $s = -1$ ="eating oats").

[7] D. Reilly, L. Cooper, C. Elbaum, *A neural model for category learning*, Biological Cybernetics, 45, 35-41, 1982.

In the first case, the algorithm receives confirmation of its classification and does not change. In the second case, it will shrink the influence radii around the considered points so that they no longer collide. The algorithm proceeds iteratively, until no collisions remain between zones associated with distinct responses.

It is important to understand that this type of algorithm will not say anything about points that are too far from the training data. In the worst-case scenario, it is even possible that the number of influence regions is equal to the number of training data. This occurs when the starting radius is too small, or when the categories are so interwoven with each other, in such a complex way, that the RCE method fails to generalize its conclusions beyond a very narrow scope. In both cases, the result is an inability, for the resulting neural network, to categorize entries that are at all distant from the training data. In other words, the network is incapable of adapting to new situations and will perform very poorly.

To avoid this effect, which is called overfitting, it may be wise to strategically increase or decrease the number of layers and the number of neurons per layer. By increasing the number of intermediate layers, and the size of these layers, one allows for the construction of non-linear boundaries between categories, which is particularly useful when dealing with categories that are tightly nested within each other. However, if the precision is too fine, the boundaries constructed will not conform to the actual boundaries of the problem.

I would like to offer an analogy to this problem inspired by numerical analysis. When one has points in space and seeks to connect them with a polynomial of the form $a_0 + a_1x + a_2x^2 + ...$ (where $a_0, a_1, ... \in \mathbb{R}$), the question arises as to what degree of polynomial one should use. Should we settle for a straight line $(a_0 + a_1x)$? A parabola $(a_0 + a_2x^2)$? A cubic curve $(a_0+a_2x^2+a_3x^3)$? One might naively think that the higher the degree of the polynomial, the greater the precision, but this is not always the case. The graph below illustrates what happens when one takes a polynomial degree that is either too small or too large to connect the points properly. The Runge phenomenon refers to the extreme oscillations that appear when the polynomial degree is excessively increased.

The number of synaptic coefficients in a neural network is, in our analogy, equivalent to the degree of the interpolation polynomial. If the network has too many synaptic connexions, the boundaries drawn by a neural network may not correctly conform to the expected behavior. In the specific case of the RCE, this will result in influence zones that are far too small, unable to generalize the known responses to new cases. In the worst-case scenario, the performance can be so severely impacted that the network will be incapable of offering any meaningful responses beyond the training data. However, it is important to note that, for the same number of synapses, it is better to have a network with multiple layers than with a single layer.

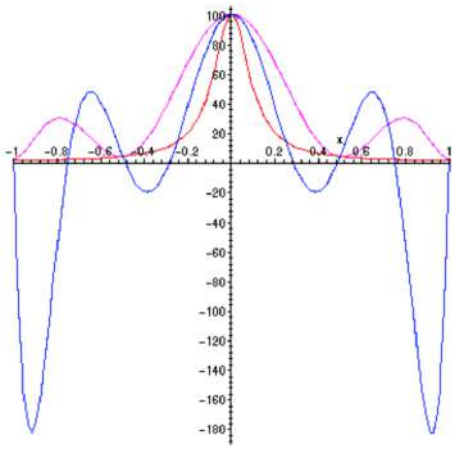


Figure 7.11: Illustration of the Runge phenomenon. The red line is the graph of the function $f(x) = \frac{1}{x^2+0.01}$. The blue lines are the interpolations of a few selected points of the function f. It is observed that higher-degree polynomials exhibit a greater deviation from the red curve (Bibm@th.net).

For this reason, the RCE algorithm will be endowed with a process to progressively increase the number of layers, and then to prune the resulting network of unnecessary neurons, iteratively, in the hope of optimizing the radii of the influence zones. Pruning of the neurons can be done during training, but it can also occur afterward. In either case, pruning requires determining which neurons are most useful to the network and which ones can be removed without much difficulty. To do this, one can, for example, remove neurons one by one and compare what effect this has on the network's outputs. Does the mean error increase by a significant margin? This comparison can be made by removing one neuron at a time or several neurons at once. The aim is to retain the sub-network (that is, a partition of the initial network) whose error function remains closest to that of the initial network (and does not exceed certain bounds concerning performance loss). More advanced methods exist, such as OBD (Optimal Brain Damage[8]), OBS (Optimal Brain Surgeon[9]) or SSM (Statistical Stepwise Method[10]).

[8] Yann Le Cun, John S. Denker and 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).

7.3.1 Genetic Programming

We have described a way to evolve the code as it learns. However, there are other approaches worth mentioning. One famous method, called genetic programming, can be used to evolve both a neural network and a top-down algorithm. This method was introduced in 1958 by Friedberg[11], before receiving refinements from Cramer and Koza. It is now being considered, in particular, for the development of artificial moral agents[12].

As its name suggests, genetic programming borrows its nobility from Darwin's theory. It involves generating a population of random algorithms and keeping only those that perform best in solving the problem at hand. At the end of this competition, called a tournament, the algorithms with the best performance are selected. Their parameters are then mixed randomly to create a new population of competitors. By repeating this process iteratively, the goal is to converge towards increasingly robust code. In biological terms, we might say that we aim to enhance the adaptation of the code with each new generation.

Suppose we want to create an artificial intelligence for the game of chess. We first need to start with an initial code. This could be top-down instructions, a neural network, or a brute-force approach. The key is that some parameters of this code must be subject to evolution. These parameters may be dedicated to evaluating positions, such as the value associated with each type of piece, the importance placed on the safety of the king, the development of pieces, the control of the board, or the pawn structure; but they could also be parameters related to the code itself, such as the number and size of the layers in the neural network (if it is a neural network) or the depth of analysis in terms of number of moves.

Each of these parameters must be associated with a reasonable probability function, which assigns a random value to the parameter within the interval $[a, b]$. This function should be general enough not to exclude interesting possibilities, but it should be constrained somewhat to accelerate the convergence process. For example, the importance given to the value of each piece might be taken from $[0, 20]$, with an infinite fixed value for the king[13]. It is important to allow null values, as it may happen that, over the course of iterations, the algorithm converges to a solution where certain parameters are unnecessary. This occurs, for instance, when the value associated with each piece tends towards the same value, or when the values of the pieces

[11] Friedberg, R.M. (1958) A Learning Machine: Part I. IBM Journal of Research and Development, 2, 2-13.

[12] Muntean & Howard, 2014

[13] In most programming languages, it is possible to assign a value called "infinite" to numeric variables. In practice, this is the highest value that can be taken by an integer. In Java, this value is 2,147,483,647. If one adds one to this value, it rolls over to zero. There is a convention to distinguish when the integer represents the largest value or when it represents strictly infinity.

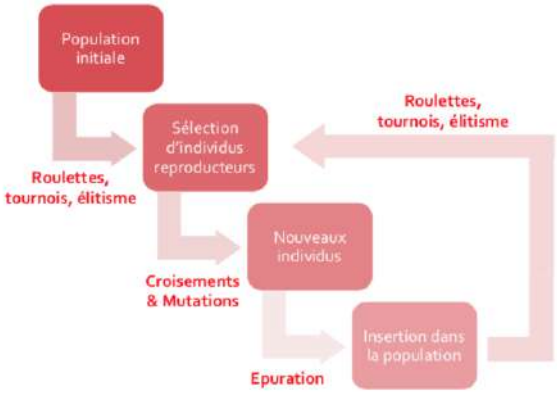


Figure 7.12: Genetic programming is an iterative process that optimizes, through selection-elimination, a population of random parameters.

hardly matter in the final evaluation of a position. Whenever this happens, the parameter can be safely removed.

Once the parameters are set, a large number of games are organised to test the algorithms against each other. A point system must be developed to determine the most performant algorithms. In chess, one point is typically awarded for a win and half a point for a draw. There are also systems to break ties (Buchholz, truncated Buchholz, and cumulative sum[14]). This system allows us to list the algorithms that performed best. The top performers are retained, for example, 10% of the best, or 5%. This percentage must be high enough to allow a certain diversity in the profile of the winners. It is indeed possible that there is more than one winning strategy: certain parameter sets, among the winners, will be noticeably different.

Finally, the parameters of the winning players must be randomly mixed to generate a new generation of competitors. The most naïve approach is to pair the programs randomly and, for each pairing, keep the average of their parameters. If algorithm A has a value of 0.736 for the "king's safety" parameter and algorithm B has a value of 0.112, their "descendant" will be an algorithm with a value of 0.424. However, taking the average may be unsatisfactory, as it can result in a leveling down. Parameter sets corresponding to effective strategies, though less representative among the

[14]The Buchholz score is the sum of the points scored by the opponents encountered. Buchholz encourages the player who has faced stronger opponents. Truncated Buchholz follows the same principle but removes the lowest opponent from this sum. The cumulative sum, finally, is the sum of points obtained after each round. According to this principle, points scored at the beginning of the tournament are more important than those scored later. This is an indicator designed to prevent a pernicious strategy: one that involves intentionally losing at the beginning of the tournament in order to be paired with weaker opponents.

winners, may disappear simply because their parameters are averaged with those of a more dominant strategy. This is why it might be preferable to randomly select one of the two values rather than take their average. A third, more subtle approach consists of averaging them through a sigmoid function. Each of the two parameters then serves as the input to a formal neuron. The output of this neuron is the new parameter to be used. There are, as one might expect, many ways to parametrize this formal neuron.

I hope to have shown, through this brief discourse on genetic programming, that there exist bottom-up programming methods that do not rely on neural networks. Furthermore, I would be satisfied if the reader comes away with even an approximate idea of how neural networks can be supported by genetic selection, or how neural networks, in turn, can strengthen this latter. Most of the performant codes encountered in the literature cleverly mix various approaches. These are hybridizations worthy of the authentic chimera of mythology, but whose effectiveness is far from mythical.

7.3.2 Unsupervised Learning

The methods we have seen so far involve placing the helm of the ship of statistical regression in the hands of the machine, but continuing to supervise its course. The human being sets the course to follow. It is the human who chooses the training data and how to interpret them, prescribing which behaviors to imitate and which to avoid.

In computational terms, we say that the programmer "labels" the data. Some inputs are considered "good" and others "bad." In the case of Buridan's donkey, for example, it is the human who prescribes which past choices were good, in order to encourage the formal neuron to replicate the advantageous decisions and steer clear of others. Similarly, in a chess software, the database consists of game situations (configurations of pieces on the board), some of which are deemed favourable and others unfavourable. In a program aimed at recognizing geometric shapes, to end with a slightly different example, we prescribe in advance which shapes should be identified: letters, squares, triangles, etc.

The judgement that labels the data is based on heuristics, or intuitive evaluation principles, designed by the human. The most famous of these heuristics in chess is the well-known rule of thumb, which states that a queen is worth ten points, a rook five points, a bishop three, a knight two and a half, and a pawn one point[15]. These heuristics are convenient in that they allow the value of encountered game situations to be quantified *a priori*. However, it is not always easy to agree on these heuristics; they

[15]A version of this can be found in Alan Turing's work, where the queen was assigned nine points. At that time, Turing had to calculate the algorithm's responses on paper, and it took him thirty minutes per move. See A. M. Turing, *Intelligent Machinery*, Collected Works of A.M. Turing, Mechanical Intelligence, p.127.

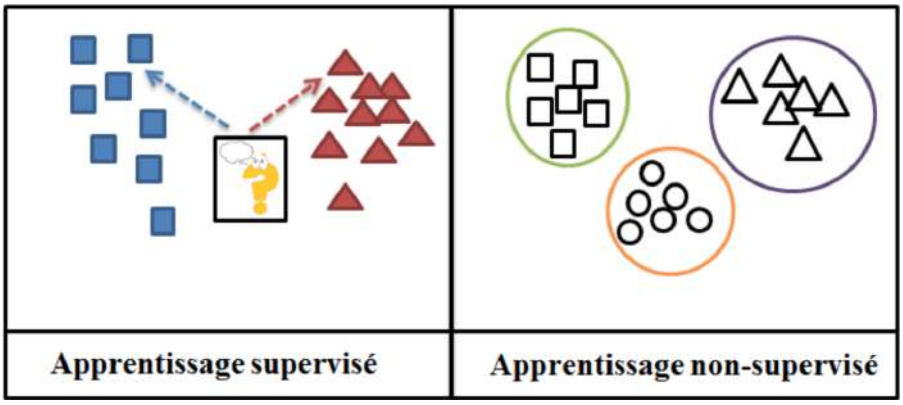


Figure 7.13: In unsupervised learning, the machine organises the data by sorting them into different classes in such a way as to maximize the statistical independence of the resulting groups.

are not valid in all contexts. Chess grandmasters' feats are full of examples where a grandmaster sacrifices material to achieve a substantial positional advantage.

This is why a new generation of neural networks, called unsupervised, operates on the principle that categories should not be pre-determined by the human designer. What do I mean by this exactly? Let us suppose we want to create an algorithm capable of recognizing handwritten digits. In the standard approach, one must give the machine a prototype example of each digit: an example of 0, 1, 2, etc. The machine relies on these examples to classify all new characters encountered. It is highly effective at identifying known digits but will not be able to create its own categories.

In an unsupervised approach, no starting prototype is given to the algorithm; we do not even tell it how many different digits the alphabet contains. We allow it to statistically compare the handwritten letters with one another to identify similarities by itself. Categorization then emerges by comparing the available inputs in pairs. This methodological difference is analogous to the one between a child learning to read in school under the wise guidance of a teacher and a child who learns to read by themselves.

The advantage of unsupervised learning is that it allows the machine to create data classes that would not have occurred to the human. Faced with an unknown alphabet, for instance, the machine can independently identify new letters. The downside, however, is that variations of the same letter may be catalogued in different classes. In the worst-case scenario, one can imagine that the machine distinguishes as many letters as there are handwriting styles: the cursive A will be distinguished from the script A, and so on.

The most commonly used categorization method is clustering, which

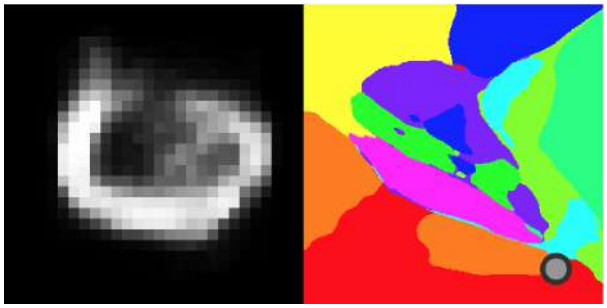


Figure 7.14: This unsupervised algorithm's task is to recognize handwritten digits without being given a prototype of each existing digit. Good performance is observed, although some letters close to 0 are classified as 6 (source: https://n8python.github.io/mnistLatentSpace/).

involves sorting the available data into similarity classes. This can be done by choosing the partition that maximizes the statistical independence of the classes from one another. This statistical independence can be measured using several indicators: Kullback-Leibler divergence, PCA, IPC, etc. I will not delve into these methods here, but I would like the reader to understand the autonomy granted to the machine in unsupervised learning.

It may happen that the obtained classes are based on statistical differences that have no meaning in human language. A machine designed for character recognition might, for example, take into account the colour of the characters, variations in orientation, or the brightness of the page—concepts that the human mind disregards when reading. Since the programmer does not know in advance how the algorithm will behave, this can result in a traceability issue: it becomes increasingly difficult to say why and how the neural network arrived at one conclusion or another. This inherent opacity is probably the most significant disadvantage of bottom-up methods.

Finally, it is worth noting that there are semi-supervised learning strategies where only some data are labeled. This ensures that the desired classes appear at the end of the process, while still allowing the machine to build its own schemas.

7.4 Epistemological Critique: Between Instinct and Reasoning

It is time to synthesize the results obtained from the bottom-up programming paradigm and highlight some of its epistemological weaknesses. Our brief trajectory through neural networks has allowed us to identify two general strategies in Deep Learning: supervised learning and unsupervised learning. In the former case, the categories into which the machine must

classify the data are defined *a priori* by the programmer. The latter stipulates, for example, that the images to be analysed represent geometric shapes such as "circle", "triangle", or "square". Furthermore, the programmer provides typical examples of circles, triangles, and squares that the algorithm can rely on to make its decision.

In unsupervised learning, by contrast, the neural network is presented with raw data and must, on its own, create concepts. These concepts emerge by grouping the data in the sample into statistically independent classes. Faced with geometric figures, the machine might, for example, group all triangles together, but just as easily, it could sort the shapes by colour, placing red ones on one side and green ones on the other. There is no guarantee that the classification resulting from this approach will align with the concepts envisioned by the programmer. It may be that large triangles, for instance, are grouped with a different class than small triangles. In doing so, the algorithm does not produce a rational criterion to identify a triangle or square based on their definitions. It does nothing but bring together objects that share certain similarities, in accordance with the adage that "birds of a feather flock together". Sometimes, the classification is even based on similarities that are not apparent to a human being.

One of the fundamental problems with neural networks, therefore, is creating a link between the statistical concepts generated by Deep Learning and the rational concepts defined using prior logical criteria, such as "having four right angles" or "having three angles whose sum is 180 degrees". This problem I am addressing requires what Plato called a "judgement with justification". It is not merely about classifying the data by placing similar objects together; one must also express, in the form of a verifiable and operational rule, the common traits shared by the elements of the same class. This rule must be general enough for all representatives of the class to satisfy it, yet exclusive enough so that the members of other classes do not.

However, as it stands, bottom-up programming lacks traceability. The answers it provides to the classification problem are not easy to explain. The programmer themselves does not always necessarily understand why the algorithm identifies a person in a photograph as either a human or a gorilla.

7.4.1 The Achilles' Heel of Neural Networks

To illustrate this difficulty, let us consider a textbook case: suppose a neural network is tasked with automating the opening of subway station doors, which was previously controlled by a human agent[16]. We would want the machine to be capable of identifying the number of people present in the station: under normal circumstances, it should open only one exit door, but

[16] I have encountered or heard this example several times in educational exhibitions, with some variations. It is probably a fictional case.

during peak hours, a second door should be allowed. To train the algorithm, we provide it with all the video footage captured by the surveillance cameras at the subway station over the past ten years. We assume that, with such a large amount of data, it will be able to determine for itself when the second door should open. After this training, we test the neural network in a real situation and observe satisfactory results. The human agent, who had been handling this laborious and unrewarding task, is then laid off — which also allows, let us admit, for some savings in the public treasury.

However, one month later, the Olympic Games begin. The subway station becomes flooded with people in the middle of the afternoon, but the algorithm fails to open the second door. People push and shove, accidents occur. An investigation is conducted, but neither the department head, nor the programmers of the algorithm, nor even the human agent who had been responsible for opening the doors, can explain the reason for this unexpected behavior of the machine. It is only then that it is discovered that the algorithm, rather than taking into account the number of people in the station, simply opens the door during peak hours, between 7 a.m. and 9 a.m., and between 4 p.m. and 6 p.m. It had observed a strong statistical correlation between these times of day and the opening of the second door over the past ten years. The statistical categories that the algorithm created to solve the problem did not use the criteria desired by its programmers! The machine had relied on a variable (the time of day) that should not have figured into the resolution of the problem, but which was nevertheless present in the training data (the surveillance videos containing this information).

What this example seeks to illustrate is that the reasoning of neural networks is not entirely traceable. We cannot precisely explain which variables the program uses, what correlations it makes, or why a particular decision is favored over another. It is precisely this lack of traceability that leads to many of the ethical issues raised by the use of bottom-up programming. I have provided some perspectives on this in Chapter , when discussing algorithmic biases. One will recall the program designed to determine whether an individual involved in criminal proceedings was likely to reoffend. A black woman might receive negative scores for minor offences, due to unwanted statistical associations being made between ethnicity and criminal records (even when ethnicity is not explicitly provided to the algorithm).

This opacity I am describing means that we cannot view neural networks as a mechanical clock, with gears visible to the naked eye, and where nothing happens without an identifiable cause. Rather, we must see them as a black box, which takes inputs and produces outputs according to unknown (or partially known) laws. Since we are unable to test the performance of this box in all possible situations, we rely on more or less reasonable assumptions. When the outputs observed during official tests of the algorithm are satisfactory, we assume that the algorithm is performing the desired function, or is sufficiently close to it. If, for example, our black box always produces the

Euclidean division a/b of the numbers a and b we submit to it, we would affirm that it is executing the Euclidean division function. However, there is no certainty that this behavior will be replicated with different pairs of numbers a and n than those used to test the algorithm.

It is important to note that the problem I am pointing out is not exclusive to bottom-up programming, but also exists, albeit in a less dramatic form, in top-down programming. In the latter, precise and binding instructions are given to the program. For example, we tell it: "open the door when such and such a criterion is observed". There is therefore traceability regarding its rules; inspecting the lines of code usually allows one to predict the behavior of the machine. Despite this traceability, we cannot be certain that the program is written correctly. We saw in Chapter 5.3 that the problem of verifying the semantics of software is indeed undecidable (by Rice's theorem); in other words, there is no automatic procedure to determine in a reasonable amount of time whether a program always does what we want it to do[17].

The essential difference between the opacity of neural networks and the undecidability of top-down algorithms' semantic properties is merely a difference in degree: while a top-down program can be verified by a human using appropriate heuristics, bottom-up programs resist verification more stubbornly. It is as though neural networks have fully embraced the idea that "algorithms are black magic". We run them and see what happens. The result is mixed outcomes. On one hand, we obtain machines that are more flexible, more adaptive, and more capable in tackling a range of new tasks. On the other hand, we pay for this with a fundamental inability to understand the internal mechanism of the algorithm and predict its responses with a high degree of certainty. It is as if, to exaggerate, we are relinquishing the use of human judgement at the outset of the process — at the initial *design* phase. This judgement can no longer intervene except downstream, when evaluating results, interpreting them, and contextualizing them.

Some artificial intelligence researchers have long argued that it is necessary to effectively combine the bottom-up and top-down approaches. Leslie Valiant, for example, began exploring hybrid architectures as early as 2008[18]. Similar strategies are undertaken by proponents of neurosymbolic programming, who combine neural network techniques with strategies typically inspired by expert systems. The reason for this enthusiasm comes in part

[17] I say "in a reasonable amount of time" because Rice's theorem only applies to algorithms with infinite storage space. In practice, this is not the case, so the theorem can be circumvented. The time required to correct the semantics of a program with a given storage base n increases exponentially with n. This means that for somewhat sophisticated programs, it quickly becomes difficult to use an automatic verification procedure that works in all cases.

[18] Leslie G. Valiant, *Knowledge Infusion: In Pursuit of Robustness in Artificial Intelligence*, Foundations of Software Technology and Theoretical Computer Science (Bangalore) 2008 (conference), January 2008.

from Daniel Kahneman's theory, the Nobel laureate in economics in 2002, which emphasizes the existence of two information processing systems in the human mind: (a) an automatic, fast, statistical system, which resembles the functioning of neural networks, and (b) a conscious, slow, rational system, more akin to top-down programming. These two systems in Kahneman's theory do not engage the same neural regions and do not follow the same principles, but both are crucial for human decision-making. In his book *Thinking, Fast and Slow*, Kahneman provides the following description[19]:

> System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration.

Throughout this thesis, I have attempted to demonstrate that there might be a third system, positioned between the first two. This System 3, which philosophy traditionally refers to as *Understanding*, is a conscious operation, though not entirely justifiable. It consists of thinking concepts in a way that is neither statistical nor rational, but intuitive. This intuition occurs through the development of a mental image, a scheme. This image is indeed inspired by concrete representations, but it points to something that transcends all possible representations. What do I mean by this? When you imagine the idea of a triangle or a tree, you are likely picturing an isosceles triangle or an oak, but the mental image you create in your mind does not refer to a particular isosceles triangle or oak tree; it refers to the abstract idea of a triangle or a tree. It is a universal archetype, a representation that aims at something beyond its representation.

It is uncertain whether such archetypes exist within bottom-up approaches; and even if they do, I believe it is promising to seriously question this issue in order to better grasp the phenomenon. The statistical regressions carried out by neural networks only think from the data themselves. All they have before them are known cases. While they can certainly produce new instances at will, it is not certain that this process of *blind generation*, which creates similar representations from existing ones, is comparable to the process of *generalization* — which seeks to understand the underlying structure common to all representable cases.

Technically speaking, I would say that the categories of the neural network are nominalistic: words such as "tree" or "triangle" are convenient linguistic short cuts to designate, all at once, the set of representable trees and triangles. Instead, the human mind refers to universals: it does not speak merely of all the known or imaginable trees, but of the tree in general. Of

[19] This quote is translated from the English by me.

course, it is not certain that a "tree in general" exists — this is a debate that dates back to the scholastics of the Middle Ages. Nevertheless, we are capable of thinking the concept of "the tree in general". This archetype is so deeply rooted in the structures of our thought that it is difficult to understand what *this oak tree* or *that maple tree* is without implicitly referring to the idea of "the tree in general". This is why the mental image, in human cognition, serves as a link between the bottom-up level (System 1) and the top-down level (System 2). It is an image that relies on the sensory world with both feet to touch, with its hands, the world of universal ideas.

I will dedicate the final chapters of my thesis to justifying the presence of this third cognitive system, properly intuitive, by considering contemporary programming paradigms. But before I embark on this thorny path, I would first like to close the critique of neural networks that I set out to produce, by highlighting some key differences between our artificial neural networks and the known functioning of the brain. This point is important for my argument, as it means that the human mind is not necessarily reduced to the operations described by artificial intelligence techniques. If indeed there is a System 3, this system must engage cerebral functions different from those we are already attempting to mimic with our neural networks.

7.4.2 The Neural Network vs. the Biological Brain

The term "neural network" might lead one to believe, as is too often the case in popular science press, that bottom-up artificial intelligence techniques fundamentally operate according to the same principles as the human brain. Yet there are at least five particularities of the neural network that distinguish it from our brain: purification, digitalization, connectivity, discretization, localization, and low integration. I will analyze these five aspects one by one, before concluding with a more general critique.

The first difference, which I call *purification*, comes from the fact that the formal neuron of McCulloch and Pitts, which I described in Chapter 7, retains only a very simplified skeleton of the biological neuron. In the mathematical model, neurotransmitters, physicochemical phenomena, or the role played by other nerve cells (glial cells, Schwann cells, Purkinje cells, astrocytes, ependymocytes, etc.) are left aside. This simplification is, I admit, inherent to any mathematical model. However, it is doubtful that the primary goal of the artificial neuron is to model the neuron in order to better understand its biological function. Rather, the goal is to draw inspiration from the biological neuron to create an effective artificial intelligence tool. Specifically, we want the neural network training process to be done *locally*, by iteratively improving the system's responses to each new available data point, rather than processing all the data with each recalibration. In this sense, the neural network seems closer to nature: it does not produce optimal responses, fully formed from the thigh of Jupiter, but rather improves

progressively, flexibly, and autonomously — through experience.

The second difference, equally evident, between the formal neuron and the biological neuron is what I call *digitalization*. The artificial neural network does not process physicochemical signals, but numbers. Rather than seeing information as a wave moving through cells, the inputs and outputs of the formal neuron are symbols: they might be a number corresponding to the intensity of a pixel, a letter of the alphabet, or the number 0 or 1. In other words, it all seems as if the neuron is doing nothing but manipulating abstract information, rather than electrical potentials, proteins, or cells.

A third difference concerns the *connectivity* of the system. As we saw, artificial neural networks present well-aligned layers. The neurons in one layer take input from the neurons in the previous layer and communicate their responses to the neurons in the next layer. This simplified structure optimizes the statistical methods implemented by these artificial neurons. However, it does not correspond at all to the nature of the human brain. In the latter, there are many different types of neurons. Some are connected only nearby, while others have connexions spanning the entire nervous system. For instance, pyramidal neurons in the hippocampus have dendrites connected to the accumbens nucleus, the reticular formation of the brainstem, or the median preoptic area in the hypothalamus.

The fourth point is what I call *discretization*. The problems processed by neural networks take a finite number of data inputs (e.g., a pixel grid) and provide a response within a finite number of possible solutions ("it's a square", "it's a triangle", "it's a circle", etc.). Moreover, it is convenient to assign one neuron to each input and one neuron to each possible response. Image analysis thus proceeds as follows: the image is first recorded by a digital sensor, in the form of a pixel grid. The value of each pixel is then input to a neuron in the initial layer of the neural network, so that each neuron represents exactly one pixel. We then observe a succession of statistical calculations, at the end of which a final neuron activates. This neuron corresponds to a possible response, such as "this is a triangle!".

If the brain functioned the same way, we would expect that the image observed by the human subject would contact their cornea, that the cornea itself would be covered with discrete receptors, that these receptors would encode each pixel of the incoming image, and transmit the information to an assigned neuron. These neurons would then interact to activate a single other neuron, corresponding to the perceived object (triangle, square, circle, etc.). There would need to be receptor neurons, each paired with the corneal sensors, as well as terminal neurons corresponding to the concepts we manipulate (triangles, squares, circles).

Yet one could contest that the images arriving at the cornea are processed by the brain discretely, by breaking down this image into a set of pixels. It is important to note that just because the cells of the cornea are small units, of types like cones or rods, which receive the incident light and give

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.

rise to vision, it does not mean that vision itself is cut into small units. It is indeed possible that these cones and rods transform an incident light wave into an electrochemical wave. This wave is a continuous process, not a discrete one. It simultaneously excites all the corneal sensors, though in a differentiated manner. These sensors, in turn, like a guitar string, emit a signal that propagates throughout the brain tissue. Neurons, in this regard, serve as relay points. They can modify the brain wave, attenuate it, or amplify it. The propagation of information in the brain may resemble what happens when a wave on the surface of a lake strikes a stone protruding from the surface. The wave is deformed by the presence of the stone (see figure ??). However, the deformation only occurs if the stone protrudes above the water's surface. Otherwise, there is no interference. Similarly, their neurons could create interference patterns in the propagation of the brain wave, depending on whether they are activated or not.

This metaphor I propose says nothing more than this: rather than representing brain processing solely in a discrete manner, seeing neurons as independent units each carrying a small quantum of information, it is conceivable to think that one could also, at least hypothetically, consider brain information in the form of a wave. This continuous representation of the brain should not, ideally, contradict the discrete representation. It would rather be a complementary model. Just as quantum physics can be described in two ways, assimilating light both to corpuscles and to a wave, I advocate the idea that the brain can also be viewed both as a digital process and as an analog process. In either case, we do not model the same aspects of the phenomenon. We do not necessarily gain the same understanding of what human cognition is.

The last difference between artificial and biological neurons, which I call *localization*, should lend even more weight to my argument. In artificial neural networks, it is as if each neuron has a role assigned by necessity. For each given process (vision, hearing, reasoning), there are initial neurons that encode the information and terminal neurons that provide a response.

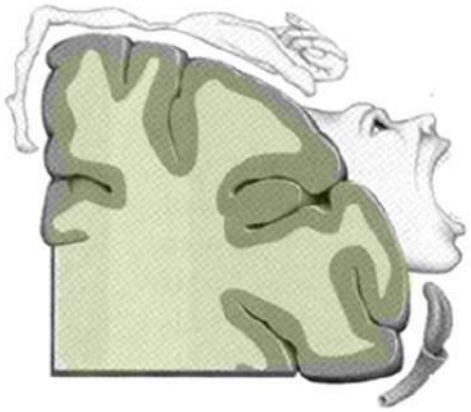


Figure 7.16: In the sensorimotor cortex, each brain area corresponds to a part of the body. This does not mean that the correspondence is point by point, but it is very clear. Some parts of the body, such as the hands, tongue, and face, are overrepresented. The volume of nerve tissue associated with these organs occupies the largest part of the somesthetic cortex.

However, in the brain, a great deal of plasticity is observed. Some neuronal regions can play different roles (vision, regulation, understanding emotions, etc.). Moreover, the role of these regions can change over time. The most striking example was observed in 2006 when a patient, after an MRI, discovered that he had lost 90% of his brain matter. Nevertheless, he was entirely capable of moving, walking, and living normally. In general, the removal of part of the brain causes the surviving cortical areas to take over the neural functions that were previously performed by the missing cells. One might even wonder if certain tasks normally allocated to the brain could, to some extent, be handled by other organs in the human body.

This phenomenon of plasticity has been known since the work of Jean-Marie Flourens (1794-1897). He performed resections of certain parts of the pigeon brain and observed that the birds were capable of quickly re-educating themselves after the operation. What seems to be predominant, behaviorally, is not so much the localization of the resection as the amount of matter removed. This powerfully suggests the idea of a certain brain plasticity, without, however, signing the death warrant of any form of localization. And for good reason: alongside this plasticity, and simultaneously with it, there are *also* spectacular cases in which neurons indeed have specific roles. One can think, in particular, of Penfield's homunculus, in the somesthetic cortex, where each cortical area selectively activates when we feel a part of our body: certain neurons are linked to the fingers, lips, ears, or knees (see figure 7.16).

I believe it is reasonable to think that if neural networks do indeed

capture a certain dimension of the functioning of the human brain, it would be naive, if not entirely reckless, to ignore the specifics of the latter. The great richness of the brain comes from its high integration. Nerve cells are highly connected, not only to each other but also to the entire organism, so much so that brain activity is correlated with the activity of the heart, intestines, muscles, etc.

Neural networks reduce this complexity to match, for each neuron, a precise discrete piece of information. It follows that the entire art of artificial intelligence today consists of reassembling *a posteriori* the links between neurons, in order to obtain a global understanding of the information each one individually carries. This is how image analysis, for example, requires linking each pixel to the others, using synaptic weights, to understand whether these pixels, taken together, constitute a known shape. In other words, each formal neuron is an isolated point, and the synapses between these neurons are binary questions that ask: are these two points linked in some way?

In contrast to this *discrete* and *localized* approach, it is possible that the human brain receives global information, in the form of a wave that stimulates all the cells of the body simultaneously, albeit in a highly differentiated manner. What do I mean by this? That the wave propagates, not *only* from one cell to another by jumping, but *also* to all the cells as a large wave. Since the wave is not uniformly shaped everywhere, differences in electric potential appear in certain regions of the cortex. It is this contrast in the electrochemical activity of neurons, more so than the intensity of the wave itself, that would allow the brain to identify known shapes. The vision of an apple, for instance, would create differences in activation between brain cells that would not be comparable to the activation differences the subject experiences when seeing a tree.

The idea that brain information arises from a differentiated activity of cells echoes, albeit analogically, Tononi's theory of integrated information. In this theory, cognitive processes such as sensation, perception, or consciousness are not assigned to precise areas of the brain, but to the ability of nerve cells to resonate and statistically differentiate themselves, through their synchronized activity, from the activity of other cells in the brain[20].

To conclude, it could be that the human mind functions holistically, starting from the global information of perception and decomposing it into independent parts corresponding to objects, colour points, each correspond-

[20]The explanation I give here is a simplified schematic. From a more technical point of view, Tononi and his colleagues define a measure of information, inspired by Shannon's work, which they call integrated information. This quantity seeks to measure whether a group of nerve cells is statistically independent from the other cells of the brain tissue. This means that we seek activation patterns, forms, within the whole that is the brain's organic activity. See notably Tononi G. et al., *Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms*, arXiv, December 30, 2022.

ing to specific patterns of differentiated activation in the brain tissue. Upon reading this conclusion, the reader should not be overly surprised. I have merely transposed the Kantian critique from Chapter 3 to the more specific workings of the brain and neural networks. In this hypothesis I propose, contemporary algorithms operate not holistically, but sequentially: they add isolated pieces of information to one another, each represented by a specific neuron, in the hope of grasping the relationships these points of information maintain with each other.

In the final part of my thesis, I now wish to gather all the critical bricks I have addressed thus far in order to reconstruct a tricameral theory of the human mind. This theory should justify the presence of a System 3 between the Systems 1 and 2 identified by the works of Daniel Kahneman. Furthermore, it should clearly show what constitutes the blind spot of artificial intelligence.

What to take away from this chapter

The second programming strategy to emerge is the *bottom-up* strategy. Unlike the top-down approach, this strategy does not seek to equip the machine with general problem-solving rules but aims to use the available data on the problem to allow the machine to learn on its own.

The most popular bottom-up approach is the neural network. This model, inspired by the biological neuron, was formalized in the 1950s by McCulloch and Pitts. It is built according to two learning approaches:

- In *supervised learning*, rational concepts are defined by the programmer. Users provide feedback to the algorithm each time it makes a mistake in handling these concepts.

- In *unsupervised learning*, the machine creates its own categories, but these are not given a rational definition; instead, they are simply statistical groupings of objects exhibiting similarities.

Despite these two strategies, neural networks struggle to abstract the general properties of the objects they are presented with. This gap suggests that there exists, in humans, a third level of thinking between the top-down and bottom-up levels. This hypothesis will be explored in the next chapter.

Chapter 8

The Blind Spot of AI: Sketch of a Holistic Theory of Judgement

I would like to summarize what has been said so far with an illustration drawn from the game of chess. This will require a brief digression. Since Deep Blue's victory over Kasparov in 1996, very few human beings have dared to challenge the machine in official matches. It did not take long, after that initial breakthrough, for artificial intelligence to triumphantly ascend the final steps separating it from invincibility. In just a few years, software like StockFish, Rybka, or Deep Fritz had become so powerful that the rare daredevils who chose to face them fell victim to utter rout[1]. The machine gradually asserted itself as an analysis engine, referee, and coach, so deep had its mastery of the game become. The statistical chances of victory for the reigning world champion against StockFish are estimated at one percent[2]. Yet chess players continue to score occasional victories, relying on trial and error. They do so by exploiting a limitation known as the horizon effect (see figure 8.1).

The horizon effect is linked to the difficulty artificial intelligence has in establishing long-term strategies when no short-term advantage is dis-

[1] Kramnik first faced Deep Fritz 7 in 2022 and, after a good start (3–1), ended with a draw (4–4). In the second match, in 2006, shortly after claiming the world championship title, he faced Deep Fritz 10 and finished with 4 draws, 2 losses, and zero wins. The match promised him one million dollars in case of victory and half that amount in case of defeat. Despite these generous rewards, no further official match has been held between the machine and a world champion.

[2] This calculation is made by comparing the highest ELO rating achieved by Magnus Carlsen over his career (2,889) with that of StockFish as of May 2021 (3,548). The 659-point difference suggests, according to the logistic scale used to compute ELO ratings, that Magnus's chances of winning are exactly 1%. It is generally acknowledged, however, that a 600-point difference at high levels of competition carries more weight than the raw numerical gap implies.

Figure 8.1: **Horizon Effect:** This position is solvable by a human but complex for a machine. White wins by bringing the king to e7, then the bishop to f8, followed by moving the knight to capture the black pawn on d7. The pawn on d6 can then be promoted.

cernible[3]. Faced with the computational complexity of a chess position, humans tend to fall back on synthetic concepts: control of the diagonals, pawn structure, etc. These concepts, acquired through a blend of experience and reasoning, are clear enough to be handled intuitively without calculating every possibility; yet they remain too complex or too vague to be expressed quantitatively in the machine's language. What value should be assigned to controlling a file? What is the worth of capturing a piece? How should a good pawn structure be defined?

Needless to say, if the machine could examine all possible moves, from the opening to the final move of the game, it would have no need to manipulate such simplifying concepts. It would have, laid before its eyes in full, the entire tree of possibilities[4]. The game would be transparent to it. In practice, however, it is impossible for an artificial intelligence to consider every possible move, as these proliferate exponentially. Assuming there are only

[3]The exact definition of the horizon effect requires some technical vocabulary. It occurs when the machine, after anticipating n moves, concludes it has found a good move, whereas, had it looked to move $n+1$, it would have realized that this is not the case. Conversely, it may judge a situation as bad after n moves, whereas a deeper search would lead to a different opinion.

[4]This expression "tree of possibilities" arises from the fact that each move can be represented within a tree. The root corresponds to the opening move, and the branches are the array of possible responses.

20 possible moves per turn[5] and that a game does not exceed 60 moves[6], we record 20^{60} possibilities, which is close to the estimated number of atoms in the universe. This clearly exceeds the capacity of today's machines.

Since the game tree in chess is far too vast to be explored in its entirety, the machine must, like the human, eliminate *a priori* the moves that appear least advantageous. This requires it to evaluate a position using concepts: piece value, control of files, pawn structure, etc. The very heart of computer science, then, lies in endowing the machine with the ability to group under a single expression game situations that may differ, by creating at least three broad categories: "winning situations", "drawn situations", and "losing situations". A situation is called winning when it offers better chances of victory than the current one; it is losing when the chances diminish. If these categories are constructed intelligently, then the algorithm, instead of stupidly analysing every possibility, can halt its analysis as soon as a sufficiently large or sufficiently small advantage is observed. This is called *pruning*.

I have suggested, in earlier chapters, that these categories can be formed in two ways. The top-down approach consists in endowing the machine with a set of analytical rules, generally called "heuristics" in computer science. These rules assign a value to each game position by looking solely at the arrangement of pieces at a given moment; they can therefore be expressed in quantifiable logical laws, examining each square, each piece, and checking whether threats are posed. The most famous example was proposed by Alan Turing, when he suggested assigning a value to each piece: 10 for the queen, 5 for the rook, 3 for the bishop, 2.5 for the knight, and 1 for the pawn[7]. This rule of thumb, well known to beginners, is often mistaken. By sacrificing a supposedly stronger piece, a clever player can gain a positional advantage. Typically, a queen sacrifice may lead to checkmate, or a bishop sacrifice may so compromise the enemy king's defence that the opponent is in turn forced to give up material. This shows that the rule of thumb alone is not enough to assess the value of a position. The fundamental question is: how many rules are needed for a satisfactory description? And more fundamentally: is it possible to reduce all the game's complexity to pure analytical rules?

The bottom-up approach, for its part, takes a very different stance.

[5]There are sixteen legal opening moves: moving a pawn one or two squares, or moving a knight. This number increases quickly after the opening and decreases toward the end of the game.

[6]Some recorded games have lasted over 300 moves. There is, however, no reason to believe this number can grow indefinitely. A game ends, according to official rules, when a position is repeated three times. As the number of such positions is finite, a game always comes to an end eventually—but it may take an enormous amount of time. That is why the machine must be equipped with synthetic concepts, allowing it to recognize when a position is blocked or looping, rather than relying solely on brute force.

[7]Note that this rule has some variants. Sometimes the queen is assigned 9 points instead of 10, and knights are occasionally given 3 points instead of 2.5.

Rather than providing ready-made analytical rules, it relies on statistics. It starts by building a vast database of so-called "solid" games, preferably played by top-level players. When it needs to evaluate a given position, it checks in the database what percentage of games featuring that position were won by Black and what percentage were won by White. This gives it a synthetic, purely empirical view of the game. Of course, the position to evaluate may not be present in its database—in which case it must fall back on analytical rules. These are used until the machine encounters a known position. If it does not reach one, it stops at a given depth (the depth being the number of moves the machine anticipates).

Thus, analytical and synthetic rules go hand in hand. The bottom-up approach, when based on self-learning techniques, does not encode analytical rules in advance; rather, it gives the machine methods for extracting effective rules from its own database. The AI thus uses a certain subset of its recorded games to deduce effective analytical rules; it uses other recorded games to test these rules' effectiveness. Once the rules have solidified, the machine can begin playing games against itself, in a manner akin to evolutionary methods, to build new databases or further refine its heuristics. In the extreme case, the machine can learn to play chess starting from no prior knowledge other than the rules of the game. This is the case of the Alpha Go program, which learns without any prior database[8]. This method of learning is called unsupervised learning, or more precisely, reinforcement learning.

Despite the machine's indisputable progress on the chessboard, artificial intelligence has not fully freed itself from the horizon effect, such that seasoned players still manage to defeat the machine now and then. A striking example of this phenomenon was demonstrated by Hikaru Nakamura, ranked second-best in the world, in a match against StockFish level 7 (3200 ELO) on March 21, 2024[9]. Instead of playing a regular match, he opted to make disadvantageous sacrifices to create a closed pawn structure (see figure 8.2), where the pieces could no longer move. Nakamura knew, however, that a knight—though considered weaker than a bishop according to the rule of thumb—would prove more powerful in such a configuration. After 271 moves, composed largely of back-and-forth maneuvers, he ended up promoting six pawns on the far end of the board—transforming them into bishops to crown his crushing victory with the flourish of an artistic checkmate.

Everything unfolds—as I have endeavoured to show in the preceding chapters—as if there existed two registers of judgement. Analytical judge-

[8]Strictly speaking, Alpha Go is the name of the Go program that defeated Lee Sedol in 2017. Leela Chess Zero, or Lc0, is the name of the chess software inspired by the methodology adopted by Alpha Go.

[9]It should be noted that, at the time of the match, StockFish was considered the strongest chess engine in the world. It had replaced some of its naive heuristics—which had cost it defeats against AlphaGo in 2017—with a robust neural network.

Figure 8.2: This figure illustrates the strategy adopted by Hikaru Nakamura (2800 ELO) to defeat the Stockfish program (3500 ELO) on the chessboard.

ment consists in making a decision based on rational, calculable concepts made of quantitative rules. Synthetic judgement, on the other hand, relies on experience, statistics, memory. Both human players and machines draw upon these two faculties when facing the board. I would like, however, to consider the idea that within the human mind, there exist concepts occupying the exact midpoint between pure synthetic judgement and pure analytical judgement. These concepts would be, in Kantian theory, the analogue of the synthetic *a priori* judgement, though not necessarily defined in the same way. They are representations that can be the object of calculation, yet cannot be fully or entirely explicated. They are ways of reducing the complexity of a game situation to a few simple, manipulable schemas, which nonetheless are reducible neither (1) to a set of quantitative rational rules, nor (2) to a set of memorized empirical statistics.

In conclusion, I hope to have suggested—relying on the example of chess—that there exist three levels of thought: (i) the synthetic level, corresponding to statistical associations (a repertoire of games linked to a probability of victory), (ii) the analytical level, corresponding to logical inferences (calculations performed to evaluate a position, using for example the rule of thumb), and (iii) an intuitive level, which resists both statistical association and quantitative logical rules. I would like to devote the next chapter to justifying the existence of this intermediate level and demonstrating its relevance. I will again draw upon the example of chess. A further chapter will then be dedicated to comparing these three levels and their respective relationships.

8.1 Intuitive Concepts

Let us return for a moment to the example of the game of chess. Unlike the machine, the human player who judges a chess position to be favourable due

to control of a row, for instance, cannot explicitly enumerate—even if they wanted to—the full set of games in which such a position led them to victory. They are merely reproducing a pattern that has proven effective; this pattern is a simplifying idea, or in other words, an abstraction, consisting in isolating one component (the open row) from the whole of the game. In making use of it, they proceed as if the row were independent of all the other factors at play in the position. They grant it, so to speak, an autonomous existence. And yet they know perfectly well that no game can be reduced to the single concept of an "open row." They are fully aware of the gap between their concept and the object to which it refers. This differentiation between words and things, between properties and objects, is the indelible mark of intuitive concepts.

Do such concepts exist in current algorithms? One might object that neural networks do nothing else when they adjust the weights in their architecture. They isolate certain parameters from the mass of data presented to them and consequently adjust the relationships between neurons. In this way, their responses to certain problems take into account the examples encountered during training, even if those examples are later removed from memory. Put differently, the structural evolution of the neural network appears capable of adapting to the cases it is presented with, without needing to recall each of those cases explicitly.

For the machine, it is important that, when constructing its neural network, information is presented as a set of symbols—typically binary numbers. The neural network processes these symbols by traversing them one by one. This presupposes, in a manner of speaking, the existence of minimal units in the structure of thought: atomic informational elements. When it comes to analysing an image, for example, these units are pixels. To enable the machine to extract categories from within this image—such as the category of an "open row" in a chess position—it suffices to present the information as a grid of numbers. Each number in the grid represents the value of a pixel (this number indicates brightness in the case of a black-and-white image; in the case of a colour image, three numbers are needed, each indicating the pixel's red, green, and blue components). Now, it is not guaranteed that the human being, when attempting to abstract a category from within an image, thinks in terms of pixels, and this for two reasons: first, this would require that the human mind break down the whole of perception into minimal units; second, it would need to be capable of enumerating each of those units one by one in order to process them and draw conclusions.

These two assumptions collide with an epistemological problem: can the human being identify pixels without first having a concept of colour? Is colour defined through pixels, or are pixels defined through colour? In the case of the machine, this interdependence between colour and pixels is of no concern. One assigns numbers and assumes the computer responsible for processing them possesses rules for manipulating them—addition, multipli-

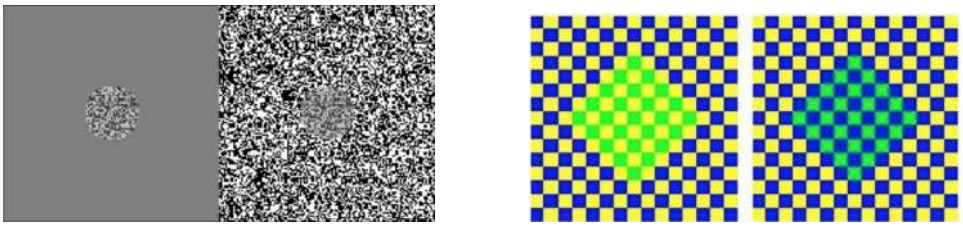


Figure 8.3: In the Chubb illusion (left), both disks are identifiable, yet the one placed on a uniform grey background appears more contrasted. In the Valois illusion (right), the green squares are identical in both images, but the presence of blue squares in the background causes them to appear more bluish to our eyes. These two experiments suggest that the perception of colour and contrast functions by isolating components through comparison.

cation, etc. But for the human being, unless one presupposes that the mind resides in the brain and that the brain is itself a calculating machine—which would make the argument circular—one observes an interference between the whole and the parts. The identification of a coloured point in a given location is influenced by the surrounding colour points, as is clearly illustrated by the Valois illusion (see figure). Even in black-and-white images, the perception of contrast is influenced by the background, as shown by the Chubb illusion (see figure).

These illusions suggest that the human mind does not reason from predefined points, but from contrasts. The segmentation of the image as a whole does indeed take place, but it does not consist in arbitrarily inscribing each point of the image into a pre-established pixel grid. Rather, it consists in isolating zones that differ from one another. This segmentation is dynamic. As soon as a form has been identified and abstracted from the image as a distinct entity, it remains possible to decompose that form in turn into smaller regions. Step by step, the human mind may arrive at points, but these are not minimal units of perception; they are temporary anchor points. This is why there is no such thing, in the concrete experience of the painter in the studio, as a pure point of colour. Each point has dimension, shape, and often harbours within itself as many contrasts as with the rest of the painting.

These remarks suggest the following: in addition to synthetic and analytic judgement, the human mind possesses a holistic approach. This consists in dividing up reality by means of differentiations. Through a shifting play of contrasts, it becomes possible to extract abstract categories from the whole of perception. These perceptual categories are that through which the human mind apprehends the sensible world. Without them, experience would be unintelligible: it would not be a painting made of distinct colours, which one could look at one after the other, but an indivisible whole. A

reality given in one fell swoop to the mind, without objects, without properties.

These perceptual categories are thus necessary for the understanding of the world and, in this sense, play a role analogous to that which Kant assigned to the categories of the Understanding. However, they do not have any existence independent from the world itself. It is just as impossible (1) to represent the world without doing so through these categories as it is (2) to represent the categories apart from the world. Point (2) is the reason why the categories in question here cannot be defined through rational and quantifiable rules. These categories make sense only because they are connected to an empirical context. Point (1), for its part, implies that the category is something other than the mere enumeration of prior sensory facts. Rather, it is what enables one to think sensory facts and to grant them an independent existence within the human mind.

This is why the intuitive concepts that the chess player manipulates while observing a game board are (1) not the mere listing of existing situations, and (2) cannot be defined by formal rules. Concepts like "open row", "strong pawn structures", or "control" are the result of an idiosyncratic segmentation of reality. They are anchor points, contrasts through which the player is able to perceive the board. Without them, the board would be an unintelligible whole. Being confronted with it would likely resemble what the beginner feels when lost amid so many pieces, squares, and threats. Incapable of seeing the game board in purely chess-like terms, the novice clings to the categories they already know: they focus on the geometric shapes of the pieces, their material, their colour, on what they evoke (a king, a knight, a queen, a soldier...). If one were to change the shape of the game, replace the classic Staunton pieces with modern or abstract ones, if the wooden board became stone, or if the game were represented flat in two dimensions, they would be disoriented; their analysis would suffer.

The experienced chess player, on the other hand, analyses the game using chess categories. They break down the board without regard for texture, colour, or shape, focusing solely on contrasts that matter for the game itself. They no longer see isolated pieces with geometric shapes; they see interdependent structures of pieces, connexions, symmetries, power relations. This is why concepts like "pawn structure" or "king safety" suddenly take on an intuitive meaning for them. This is not due to a formal definition made of rules and explicit calculations, nor to anticipating the opponent's moves. Nor is it simply because they recall past games or have identified patterns (though this plays a large part as well). It is because their way of decomposing the whole of perception (the game board as a whole) into meaningful parts (units of analysis) follows a different structure. In other words, they possess a better capacity for intuitive abstraction.

This observation explains why, when two chess players—a beginner and a master—are asked to memorize a configuration taken from a real game, the

master performs better; but when the configuration does not come from an actual game, but rather from a random arrangement of pieces on the board, it is the beginner who excels. And for good reason: their vision of the game is not structured by chess concepts, but by geometric ones. For them, the problem resembles placing shapes in certain locations. The master, by contrast, when faced with a configuration that violates all the rules of chess, experiences a disorientation akin to what the beginner felt when discovering the game for the first time. Everything seems strange, bizarre, incongruous. They can no longer easily segment their perception into intelligible units.

Of course, one might argue that the differences I have just described between the chess master and the beginner are analogous to what happens with a neural network such as AlphaGo. In the latter, categories are built up as the machine encounters new game situations, changing the way information is structured. This view is supported by the fact that neural networks today are capable of defeating any human player at chess. I have shown earlier, however, that the horizon effect still persists, revealing that these artificial intelligence categories, though more efficient on average than those of humans, seem to follow different rules. They are based on statistical regularities between the pieces, not on global contrasts. They rise from the particulars toward the general; whereas one is right to wonder whether the human mind does not reason from the whole.

I acknowledge that the very existence of a holistic judgement capacity within the human being remains purely hypothetical. I am content, at the close of this chapter, if I have succeeded in shedding light on a blind spot of computer science. It may be, in the final analysis, that the intuitive concepts I have described can be reduced to statistics. Yet I believe it would be a mistake to presume so. I would like instead to consider the possibility that intuitive concepts occupy a middle ground between a purely analytical approach, founded on rationality, and a purely statistical one, founded on experience. This is why, in the next chapter, I will argue for a model of the human mind composed of three levels: the top-down level of Rationality, the bottom-up level of Sensation, and the holistic level of Perception. I will attempt to show how this theory connects with the Kantian and Platonic theories explored in previous sections.

8.2 A Tricameral Theory of Cognition

I have suggested that intuitive concepts, if they are to represent something beyond mere statistical categories, must occupy a space between experience and rationality. For this reason, I would now like to argue that human judgement—like democracy—is composed of multiple parliamentary chambers, each of which plays a distinct role in processing information and deliberating.

The first chamber of the mind is content to legislate. It lays down logical rules and analyses them with precision. This is Rationality. The concepts it manipulates are rigorously defined: they are abstract and mathematical, quantifiable, and implementable within a computing machine. Yet its reasoning is far removed from the affairs of the world.

The second chamber dwells close to the realm of everyday experience. It makes semi-automatic decisions in the ordinary matters of life. This is Sensation. It is concerned with the practical affairs of the world. It weaves correlations. It recognises that the law, here, is not strictly defined, but only statistically so. It manipulates not well-defined rules, but implicit, practical, and often unconscious ones, which I call percepts. These are the fruit of experience. They cannot be fully described through language.

The third chamber, finally, is that of intuitive concepts, which lack rigorous definition, yet can still be the object of judgement. These true but unjustified judgements arise when a case presents itself that cannot be elucidated by laws nor resolved through effective practical procedures. This chamber, which I describe as Perception, draws at times from Rationality, at times from Sensation, striving to amend their limitations. Perception refers indeed to rational law, but applies a more flexible jurisprudence, one that takes context into account. It thus ensures dialogue between the other two modes of thought—much like the judiciary, in a democracy, whose role is to interpret the law and ensure its enforcement across all other levels of power.

These three chambers I envisage—though not without a touch of humour—clearly evoke Kant's Sensibility, Understanding, and Reason. I would now like to describe them more precisely, in order to show how they may be related to Kant's theory, and in what ways they diverge from it.

8.2.1 Rationality: Abstract Concepts

Rationality, to begin with, must be understood as the faculty for hypothetico-deductive reasoning. Rational judgements consist in the manipulation of symbols according to precise syntactic rules, akin to those I mentioned when discussing Gödel's theorems. These rational judgements are perfectly embodied in the top-down programming paradigm, and particularly in the model of expert systems.

Within Rationality, there is no immediate relationship with the real. Ideas, sensory data, lived experience, objects, and values are all represented through numbers. The essential difference between this Rationality and Plato's Reason lies in the latter's twofold structure: *dianoia* and *noésis*. The rationality I introduce dispenses with *noésis*. In the contemporary conception, there exists no intellectual intuition, no contemplation of ideas. In Frege, one finds indeed an intuition behind the axioms, but this is reduced to a feeling of self-evidence—Cartesian in its starkness. One might say that

the rise of modern rationality is characterised by a syntax divorced from semantic content, where truth is nothing more than a convention of signs. Abstract concepts themselves may be represented through mathematical symbols. In computer science, a triangle is defined by assigning numerical variables to represent, for instance, the number of sides or the sum of the angles.

Rationality concerns—cue drumroll—rational concepts. These are defined through univocal properties, most often verifiable and quantifiable. Natural numbers, for example, can be constructed from a set of foundational properties, as in the Peano axiomatic system (discussed in Chapter 5.2):

1. The element called zero, denoted 0, is a natural number.
2. Every natural number n has a unique successor, denoted $s(n)$ or S_n, which is also a natural number.
3. No natural number has 0 as its successor.
4. Two natural numbers having the same successor are equal.
5. If a set of natural numbers contains 0 and contains the successor of each of its elements, then this set is $\mathbb{N}$.

Once this set of axioms has been established, artificial intelligence consists in developing an expert system capable of automatically deducing the consequences that follow from them. This top-down approach allows, at best, the discovery of arithmetical properties that had not been anticipated by the human mind. It may also be used to complement (or simplify) proofs concerning results already known.

One example, devised by Donald Knuth in the 1960s, consists of an algorithm for computing the first n prime numbers[10], where n may take any finite value. For this algorithm to be valid, it must not merely return prime numbers whose validity the user will check *a posteriori*. For in that case, one would be leaving the realm of purely rational concepts and entering the domain of experience[11]. What the algorithm must do is apply a method that

[10] Donald E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, Addison-Wesley, 1997.

[11] Indeed, it is reasonable to think that mathematical numbers, although they are idealities, can be the object of sensory experience. A good example is provided by the Collatz sequence: take any integer; if it is even, divide it by two; if it is odd, multiply it by three and add one. Repeating this as often as needed, one observes that one always ends up returning to the number 1. Starting with 7, for example, we get 7, 22, 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, 4, 2, 1... This Collatz sequence is currently unproven, and it remains entirely possible that, for some integer, the sequence never reaches 1. Thus far, however, even the most powerful computers have failed to find a counterexample. We are therefore faced with a numerical experiment that is valid empirically, but not

is rationally proven; that is, a proof reducible to the Peano axioms. Concerning Knuth's algorithm for prime numbers, a proof may be established either by drawing upon a result published in the early 1850s by Chebyshev, or by relying on a demonstration from around 1934 by Paul Erdős[12].

Top-down approaches, despite their computational power, teach us nothing beyond what is already entailed by the axioms. Moreover, it may happen that these axioms are insufficient to perfectly describe the objects to which they refer. Gödel's incompleteness theorems, as discussed in an earlier chapter, show that the Peano axioms do not allow one to demonstrate the consistency of arithmetic. They are, in a certain sense, insufficient—and it is no small matter to circumvent this thorny difficulty.

8.2.2 Sensation: Statistical Percepts

The second faculty I introduce—or rather preserve—is Sensation. It must be understood as the faculty of establishing automatic, and generally unconscious, associations between various sensitive experiences: impressions, shapes, sounds, colours, emotions, and so forth. It is, however, important to note that there is no determined perception within sensation. One cannot say of the sensation of a triangle that it is a triangle, for in doing so, one steps outside the realm of sensation. The word "triangle" refers indeed to things not presently contained within sensation: it refers to the idea of all other triangular things, to every possible and imaginable triangle. This is why sensation cannot be the object of a linguistic definition. Nor can one take sensation as an object and assign it predicates: it is hot, it is sharp, it is dark. All these predicates, once again, point beyond what is felt in the immediate present.

Sensation exists essentially only within relations of similarity and differentiation. A sensible form is a contrast that detaches itself from the whole of sensation, acquiring a relatively autonomous existence. One can only become conscious of a sensible impression insofar as one is also conscious of what it is not. I have illustrated this on several occasions through perceptual illusions. When discussing chronostasis, I argued that there is no present instant given all at once in any objective fashion. The instant, just like the sensible object, is but a retroactive segmentation within a set of neighbouring temporal events fused together. This is why sensitive judgements, too, can only be *a posteriori* associations.

Sensitive judgements occupy a position analogous to Kant's synthetic *a posteriori* judgements. The difference lies in the fact that, for Kant, the synthetic *a posteriori* judgement allows one to relate concepts to each other,

rationally. This suggests that formal proofs belong to a different order than experimental verification.

[12] These demonstrations concern Bertrand's postulate, which states that there is always a prime number between any two numbers k and $2k$ (where k is any positive integer).

as when one says "this table is made of wood." It is indeed *this* particular table, and not all tables, yet there is nonetheless a use of linguistic relations. In my conception of sensitive judgement, by contrast, the associations are subconscious. One may indeed believe that this table is made of wood, but one would be incapable of justifying or explaining it. The objects we handle are not completed perceptions, but mere percepts[13].

I shall hypothesise that percepts, within the human mind, are defined extensionally. What I mean by this is that the percept of a tree, for instance, is nothing more than the (unconscious) set of trees the human mind has encountered or imagined. Within this category of percept, there exists no process of abstraction consisting in extracting, from all known and observable trees, some essential properties—such as having a trunk, being composed of bark, roots, branches, leaves, and so on. To put it differently, the percept of a tree is not accompanied by a general schema, by a vague or caricatured idea, by a symbol of a tree meant to encapsulate, in itself, all the particular objects it designates.

Might this be the reason why the percept is broader, more fluid, than the rational concept? It may indeed happen that one defines trees as having leaves, only to then discover a tree without any. If one adheres strictly to the rational concept, this leafless tree is not a tree; but from the standpoint of the percept, this trunk without leaves so strongly resembles its leafy cousin that one will be inclined to pair them together. It will then fall to us to extend the rational concept of tree by distinguishing, within it, the family of deciduous trees (which lose their leaves in winter), semi-deciduous trees (which shed their leaves as new ones appear), and non-deciduous trees (which do not lose their leaves).

Neural networks do not manipulate concepts, but only statistical percepts. They assign specific weights to the synapses of the network so as to discriminate a table from a chair, or wood from stone. However, they do not provide any linguistic definition of these concepts. They simply state: "this data you give me most resembles that one." They thus associate new information with already known data classes. Moreover, each new pairing they carry out is liable to alter their structure. When a link is established between data A and data B, something is learned about A, but potentially also about B. They may therefore evolve over time.

A software such as ChatGPT, though absolutely brilliant in its command of language, sorely lacks abstract judgement. Unless provided with additional rules, it cannot easily grasp the concepts it manipulates. It sometimes fails to construct coherent stories, to play chess meaningfully, to identify its sources or to criticise itself, contenting itself with repeating the most probable sequences of words[14].

[13] Deleuze: "The philosopher's profession is to create concepts, the artist's is to create percepts." (Letter I, Fifth Idea)

[14] It should be noted, however, that version o1 of ChatGPT, released in September

A neural network can be trained to manipulate concepts, even if it has no representation of them. If given sufficient data, it will be capable of performing arithmetic operations and determining empirical properties by observing correlations among certain sequences of numbers. For instance, Machine Learning methods have been employed to guide human intuition regarding the Birch and Swinnerton-Dyer conjecture. This conjecture describes the set of rational solutions to the equations defining an elliptic curve. No proof has yet been discovered for this problem, though it is one of the seven Millennium Problems—each carrying a one-million-dollar prize from the Clay Institute. To represent this problem using a neural network is to decide, rather than analysing the concepts, to treat the problem as a black box linking inputs and outputs. From such correlations, useful conclusions may emerge. We have moved from an essentialist approach to a probabilistic one.

It is important to note that the statistical associations made by a neural network are not identical to human percepts. The machine structures data by slicing reality numerically. It segments a photo into pixels, an audio recording into time units, it discretizes space so as to associate, to each point, values representing temperature, brightness, albedo, and so forth. The human being, by contrast, possesses no minimal units of perception. One begins with the whole of sensation and carves it up so as to obtain primitive forms, which exist only in the contrasting relation these forms maintain with the entirety of perception. Human sensitive judgement is thus not the pure analogue of the machine's synthetic judgement. There are, however, enough commonalities for some to claim—through a kind of mythological metonymy—that the machine is capable of perception.

8.2.3 Perception: Intuitive Concepts

The third level, caught in the vice between the two preceding ones, is that of Perception. This faculty brings together the diversity of sensations under a conceptual unity that I shall call the *intuitive concept*. The difference between this intuitive concept and the simple percepts of sensation lies in the fact that the intuitive concept is transcendent, whereas the percept is immanent. What do I mean by this? That there is always, in the intuitive concept of a "table", something more than what can be found in any particular table. The concept of table, like the concept of triangle, is not this table here, nor that triangle there. It is a general form, of which concrete instances are merely incomplete or imperfect representatives. The intuitive concept is thus akin to what Plato called the Idea of triangle, of tide, of

2024, has improved in these respects. It is now capable of creating a coherent narrative thread from beginning to end. There remains, alas, a fundamental architecture that seems inflexible. The stories produced by the algorithm could bear the title of a book by Jean Ségui: "Always the Same Story."

howling wind, or of mud. It is a unity that one represents beyond sensation, and in which sensible objects participate. The difference from Plato lies in the fact that these intuitive concepts do not exist beyond the things they designate. There is thus an epistemological tension between empiricism and idealism: on the one hand, the concept of triangle does not exist; it merely designates all possible triangles in the world. But on the other hand, it is equally true to say that the intuitive concept is not, in itself, any of the sensible triangles. It is an idealised generalisation.

It is important to note, however—and this marks another difference from Plato's theory (see chapters 1.1 and 1.2)—that this idealised generalisation of the triangle, as we find it in the intuitive concept, is not a rational idea either. Rational ideas, indeed, function by means of symbolic definitions: to say of a triangle that it is "a planar figure composed of three sides". Yet when we manipulate the triangle in our intuition, we do not necessarily possess its rational definition. Piaget's studies indicate that a child may be able to represent an object very quickly, but it takes more time for them to give a rational structure to the objects they manipulate. Thus the child possesses the concept of object from the sensori-motor stage, only a few months after birth. They are capable not only of drawing a first distinction between themselves and the world around them, but also of understanding that objects possess a certain permanence. If their rattle is hidden behind an obstacle, they will go around it to retrieve it. The object does not vanish with the perception they have of it. It exists beyond the sensible context in which it first appeared. It is thus more than a simple contrast between part and whole; it is a reality unified by a representation.

I therefore hold that the mere act of speaking of a "table" or of "wood" already belongs to another order of judgement than those we have previously encountered: the conceptual judgement. Perception is an attempt to rehabilitate Kant's *Understanding*, that great *oblivion* of contemporary philosophy, but without recourse to the *a priori* forms of intuition and the categories. Does this mean that I am renouncing altogether the idea that there exist innate structures in thought? Far from it. I seek, however, to make two distinctions.

The first—and perhaps the easiest to grasp—is that the *a priori* forms I shall defend are not *a priori* in the most universal sense. They were not born with humanity since the dawn of time, by the will of a divine creator. On the contrary, they were shaped through the evolution of species, at times showing considerable variation from one culture to another. While I am critical of artificial intelligence's pretensions to capture the properties of the human mind, I do not deny that these properties are sculpted by experience.

The second difference between my concept of intuition and Kant's is more essential: it lies in the fact that the intuition I aim to defend does not rest upon the Euclidean theory of space or the Newtonian line of time. I have sought to show, through the analysis of Kantian intuition, that there

are properties yet more fundamental than time and space. Time and space, indeed, appear as the totality of sensibility, the receptacle within which there could be no objects nor events. It is, however, doubtful that this form constitutes an *a priori* in the strictest sense of the term. Our intuition has been shaped by our experience of the world. Our representation of space has been fashioned in the manner of Poincaré and Lobachevsky, through an accumulation of sensations and sensitive experiences.

What is more fundamental than space and time, and stands, as it were, upstream of them, is the fact that the minimal units of perception—such as points in space and instants in time—exist only through the Whole of that perception. Points exist only in a neighbourhood of points bound to one another, and instants exist only in a neighbourhood of interdependent temporal events. I would thus like to substitute for these pure intuitions of space and time a holistic intuition common to all spatial and temporal phenomena. This intuition exhibits a structure. An object cannot be thought outside of a space that encompasses it. That space itself is composed of the other objects. There is thus an interdependence between points and space, between the whole and its parts.

To this first form of intuition, purely holistic, I would like to add a second intuition, which is transcendent. The use of the word *holistic* refers to the capacity to apprehend the whole; the use of the word *transcendent* refers to the capacity to conceive what lies beyond the whole as it can be represented. What does this mean? I said that when we think of a "triangle", we conceive more than this triangle here or that triangle there. The triangle, in its concept, is neither an obtuse triangle, nor an acute triangle, nor a right triangle: it is what is common to all triangles, possible and imaginable.

There is, all in all, something extraordinary, something almost magical, in this faculty we have to detach ideas from the objects they designate in the world proper, as if these ideas existed prior to the world itself. Descartes was quite right to insist that there are ideas within us which exceed our capacity to represent them. He gave as an example the idea of the infinite. This idea is clearly not something we could have invented on our own. Our minds are barely capable of memorising a dozen digits (from 0 to 9), which we combine rather slowly. Only Chuck Norris, if one believes the joke, ever counted to infinity... twice. But if someone were to boast of such a feat to me, I would reply with this line from James Joyce: "Are you a god, a jackass, or a blithering ass?" Be that as it may, when we contemplate the horizon, when the sea unfolds before our eyes, when the stars look back at us, we see no limit to the expanse of space and time. Whatever boundary one might try to imagine, whatever impassable border one might trace on the map of the universe, a single leap of the imagination suffices for the world to expand.

I shall now specify more precisely the nature and limitations of these holistic and transcendent intuitions. But before doing so, I would like to

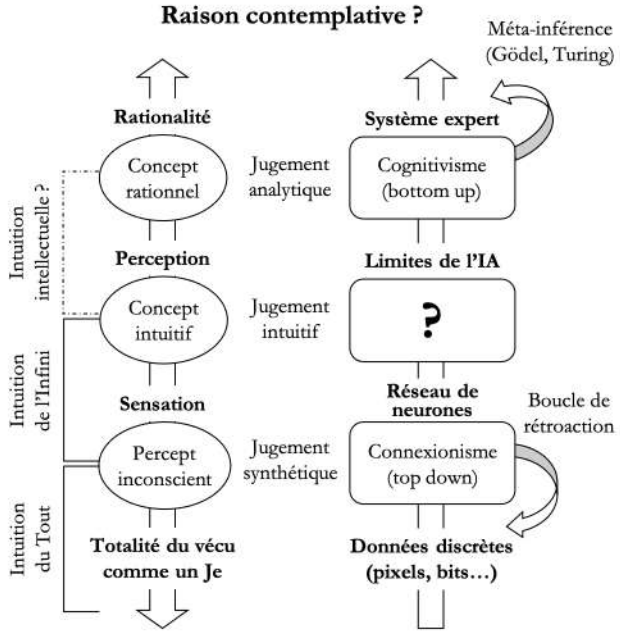


Figure 8.4

sketch a general outline. Figure 8.4 summarises the tricameral theory as I defend it. One observes that the human mind, represented on the left side of the diagram, is endowed with three faculties. These faculties emerge from the Cartesian "*I think*", that primordial state of perception in which the experience of being something rather than nothing is given as a totality. In that state, there is no distinction between subject and object, between thought, sensation, lived experience, and the external world. What is, is only what is sensed, thought, and lived in the present moment.

The decomposition of the totality of lived experience into individual objects does not occur all at once; it is a process carried out through successive comings and goings between the three faculties of Sensation (unconscious percepts), Perception (intuitive concepts), and Rationality (rational concepts). These levels are mutually supportive: objects exist through the whole, and the whole exists through the objects. The act of moving from a lower level of cognition to a higher one—for example, transforming a percept into a concept—is precisely what I have called *judgement*.

There are thus three types of judgement. Those that consist in passing from the totality of lived experience (the *I think*) to unconscious percepts (Sensibility) are synthetic judgements and rely on the intuition of the Whole. Judgements that allow percepts (Sensibility) to be gathered into intuitive concepts (Perception) are intuitive judgements and rely on the intuition of the Infinite. Finally, judgements that allow us to abstract the essential

properties of intuitive concepts (Perception) in order to manipulate them through logical-deductive reasoning (Reason) are called analytical, and they rely on a hypothetical intellectual intuition (affirmed by Plato, but contested by Kant, for whom speculative Reason has lost all contact with the sensible world of experience).

The machine, represented on the right side of the diagram, functions differently from the human mind. It does not originate from an original *I think*, but from a pre-established set of discrete, independent units (pixels, points, numbers). Algorithms are tasked with associating these discrete units to reconstruct a global vision through a purely ascending movement (from parts towards the whole). The machine thus possesses only two levels of processing: (1) that of neural networks, which corresponds in the human being to the level of Sensation (bottom-up programming paradigm), and (2) that of expert systems, corresponding to Rationality (top-down programming paradigm).

The analyses I have offered throughout this book aim to shed light on two possible gaps in contemporary programming paradigms. These gaps may be of two kinds: (i) the absence of the level of Perception (corresponding to *Understanding* in Kant), and (ii) the absence of metacognition (corresponding to Plato's contemplative Reason), which explains why algorithms are limited by Gödel's theorems and the Halting Problem (Turing).

Before concluding this part of the thesis, I would like to characterise the two intuitions in question: the intuition of the whole and the intuition of the infinite. For it is crucial to isolate the ingredients that set them apart from the machine's mode of functioning.

8.3 Critique of Sequentiality: The Intuition of the Whole

I have sketched the three dimensions of the human mind—Sensation, Perception, and Rationality—but I have thus far remained silent on how Kantian intuitions of space and time ought to be revised. Rather than confronting this issue head-on, I would prefer to highlight two ingredients particular to contemporary algorithms that are, quite precisely, absent from human intuition. These ingredients are sequentiality and univocity. I hope that by laying them bare on the epistemological operating table, the reader will better understand why machines are not—or not yet—capable of becoming true Frankensteins.

Contemporary computers operate sequentially, defining a series of operations to be executed one after the other. Only once the first is completed can the machine proceed to the second, and so on, recursively. Even so-called parallel programming techniques, based on the use of multiple threads—to borrow the jargon of programmers—do not escape this law. They merely

juxtapose different series of operations. In this respect, the machine can be compared to a long chain, composed of indivisible links, each holding the others in place. Unless it is read by a deterministic Turing machine, one that would in practice fall under the domain of quantum physics, it is not possible to perform more than a single step per unit of time.

Now, when one turns to the workings of the human mind, it is far from evident that thought unfolds in such a sequential fashion. On the contrary, it would seem that our awareness of a given instant t depends, to some extent, on our awareness of the next. It is only once that moment has passed that we can, retrospectively, take it as an object of thought—so that the thinking process unfolds in a continuous back-and-forth between "what I am thinking now" and "what I think I was just thinking a moment ago."

Perhaps this is why our perception of time is riddled with reversals? When we hear someone walking down a hallway, their shoes striking the tiled floor in regular rhythm, we tell ourselves: "someone is walking and making noise." Yet in the flow of our consciousness, we first heard the sound of the shoes, and only then did we infer that someone was walking. Our initial reasoning, then, was: "there is noise, someone must be walking." This latter reasoning is not *stricto sensu* a logical deduction. It is rather an abduction—or, in other words, a search for the best hypothesis to explain the existence of the sound. In our example, we assume that it is the walking person who is its source—but this is no logical necessity. In truth, this search for a cause suggests that we cannot become aware of our own process of thought without placing the various elements of that process (the sound, the person walking...) into logical or causal associations. The intelligibility of any given instant depends upon the intelligibility of the whole.

This phenomenon, which I have just described in relation to time, also has a counterpart in space. The perception of a scene populated by geometric objects, colours, textures, involves not only the eyes scanning from one object to another, but also an understanding of the whole. What I mean is that the global arrangement of the scene—that is, the relationships the objects maintain with one another—influences our perception of the individual elements. This fact is particularly well illustrated by the Ebbinghaus illusion, in which two circles of identical size are shown in two different contexts. In the first, the circle is surrounded by larger circles; in the second, by smaller ones. One observes that this affects the perceived size of the original circle. In other words, the perception of an isolated object depends on the presence of other adjacent objects.

From all these arguments, it follows that consciousness—far from operating in a purely linear fashion—involves patterns of association that require constant movement between the present moment, the past (memory), the future (prediction), or between cause and consequence. More rigorously, we might say that the notions of past, present, future, cause, and consequence are the result of flattening the flow of thought; they arise when one attempts

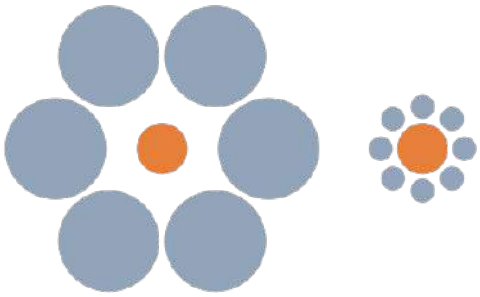


Figure 8.5: The Ebbinghaus illusion illustrates how a circle surrounded by larger circles will itself appear larger.

to structure the ideas traversing the mind. This does not render them false. They are, so to speak, *a priori* conditions of possibility—a kind of skeleton without which it would be impossible to objectify our thinking process.

8.4 Critique of Univocity: The Intuition of the Infinite

Another property of computers that deserves to be placed under the razor's edge is univocity. By "univocity," I mean a language that deliberately leaves no room for ambiguity. Programming is founded on logical, clear, and indelible instructions. When an instruction is associated with a variable, as in the expression "$a = b$," it means that a takes the value of b. This relation is non-commutative. In a programming language such as Java or C++, "$a = b$" does not carry the same meaning as "$b = a$." In the first case, a is a variable and b is a value; in the second, b is a variable and a a value. To offer a concrete example: a variable might represent the temperature of a room, the name of your doctor, or the amount of money in your bank account. A value, meanwhile, is a specific number of degrees Celsius, a name, or a given sum of pounds. Thus one may write "$T = 26°$", but one would not write "$26° = T$". In other words, a temperature can be equal to a certain number of degrees, but the degrees themselves are not, in and of themselves, a temperature.

This property that I am describing stems from the very definition of Turing machines. The latter are composed of an infinite— or rather indefinite—tape, which may be prolonged indefinitely. At each position along this tape, no more than one unique symbol can be inscribed. This symbol must be selected from a set Σ. The canonical example is to limit the symbols to 0 and 1, but in truth, the choice of symbols has no bearing on the resultant properties of the machine nor on its expressive power. It can do exactly the same things with two symbols or with a million, though it will perform

more rapidly in the latter case[15]. Be that as it may, the set Σ denotes the set of values. In our temperature example, it comprises all possible degrees Celsius, from -273.16 degrees up to positive infinity. In the bank account example, it ranges from negative to positive infinity. In the case of the doctor's name, it consists of the letters of the alphabet (with several cells on the tape required to spell out the full name).

It becomes clear that there exists, *a priori*, a hierarchy in computer language between the values Σ and the cells of the tape (the variables). This means that, in computational language, there is a pre-established difference between the subject (the variable) and the predicate (the value). One could say that the subject-predicate relation is not reversible. From the outset, what is being spoken of is posited as a subject upon which well-defined attributes are pinned.

Now, it is far from evident that such a hierarchy exists within the human mind. When we say of a person that they are haughty or miserly, we sometimes learn just as much about what it means to be haughty or miserly as we do about the person themselves. It is even possible to illuminate the nature of the predicate "miserly" or "haughty" by looking at the subjects to which these words are applied. Everything happens as though, on the one hand, attributes possessed an independent meaning apart from the persons they describe; and on the other hand, as though only the existence of miserly and haughty individuals gives these words their intrinsic intelligibility. Put differently, there is a relationship of interdependence between subject and predicate. One may reverse the relation "Croesus is a miser" into "Misers are like Croesus."

This commutativity between subject and predicate, particular to human language, is not without its difficulties. It implies that there exist no elementary and foundational variables, no minimal units of meaning on which to erect the entire edifice. Let me be more precise. In the Turing machine, the symbols Σ form an irreducible set. These characters are assumed to have intrinsic, autonomous meaning, independent of their use in the program. Even if a symbol a is never used, that does not mean a is undefined. In human language, by contrast, things are rather different. If you have never heard of a word, and it does not appear in a dictionary, you cannot know what it means. We understand French words only by referring to the contexts in which they appear. That is why dictionaries are full of references to classical texts. One systematically tracks the usage of this or that term of the French language in the hope of illuminating its various uses, its nuances, and its evolutions as well.

[15]This property derives from the fact that any symbol may be encoded in binary form. To represent a positive integer, for example, it suffices to switch to base two. The number 9 is thus written as 1001, requiring four cells on the tape. This increase in the space required explains why the machine will execute more slowly: it must travel a greater distance along the tape before reaching the end of its process.

If what I have just said is correct, it follows that one cannot suppose that the words of the French language (or any other language, for that matter) are defined from a small set of axiom-words from which all others may be derived. Such a system must presuppose that certain words are known in advance, that their definition does not depend on the way they are used to define other words. Now, this is not necessarily how the human mind proceeds. For us, semantics appears to be highly self-referential. We understand the terms of language only by weaving them into the definitions of other words, which in turn refer to other words, which in turn may depend on the original ones.

It may be objected that the Turing machine, if it were to use words from natural language, need not rely on axiom-words; it may make do with letters. The set Σ is then composed of the 26 letters of the Roman alphabet. When it does so, however, it is not manipulating words in terms of their semantics and definitions, but only their syntactic form. When it comes to writing the definitions of these words, it must indeed assume that some words are known in advance, that they can be used to define the others. The use of letters of the alphabet thus only circumvents the problem.

This is, moreover, the reason why philosophers such as John Searle believe that a machine does not understand the semantics it manipulates. It merely uses letters devoid of meaning. To his mind, it makes no difference whether it manipulates the Roman alphabet or the Chinese one. In either case, it is simply applying grammatical rules to construct words it does not comprehend. His argument is popularised by a thought experiment known as "the Chinese Room". He imagines a human being shut in a sealed room, tasked with communicating via written messages with someone outside the room. He does not understand the message, but possesses a voluminous table that tells him exactly which symbols to write in response to each message he might receive. For the sake of realism, I suggest that this man be called ChatGPT. John Searle says of this chatbot:

> The point of the argument is this: if the man in the room does not understand Chinese on the basis of implementing the appropriate program for understanding Chinese then neither does any other digital computer solely on that basis because no computer, qua computer, has anything the man does not have.

To summarise what I have said so far, it would appear that the human being, when reasoning, grasps the intuitive meaning of the words they manipulate, even though they are incapable of defining them without generating logical circularities. The machine, for its part, cannot tolerate circularity, yet it does not necessarily seem to understand the semantics of the terms it handles.

I would like to illustrate this high degree of self-referentiality inherent in human language through a concrete example known as "the dictionary

paradox." Let us suppose that a child, free of all preconceptions, sets out to look up the meaning of a complicated word in a dictionary more voluminous than an almanac. Let us assume that the book is vast enough to encompass the entire vocabulary of philosophy and neuroscience, and that all definitions within it are perfectly consensual. Let us also suppose that our curious child possesses only a very limited vocabulary, but has an innate capacity to learn new words thanks to an exceptional memory. They are familiar with grammar and will never encounter syntactic difficulties, though the theories underlying the words they encounter would entirely escape them.

It is likely that, upon encountering the definition of "consciousness," our knowledge-hungry child—closer, in this instance, to a word-processing machine than to a human being of flesh and blood—would stumble across numerous unknown terms. The only thing they could say of these increasingly mysterious predicates is that they are all, taken together, equivalent to consciousness. Should they, for instance, be so unfortunate as to fall upon a phenomenological dictionary, and read that consciousness is a "qualitative flux of inner states," they would ultimately learn just as much about consciousness as they would about the "qualitative flux of inner states"—namely, not much of any substance.

To pursue their inquiry further and attempt to loosen consciousness from this monolithic identity—assuming, of course, that they have not already been discouraged by the phenomenological jargon—they would need to understand what is meant by a "qualitative flux" or an "inner state" by referring to their respective definitions. Alas, in doing so, they would likely encounter yet more phrasings and terminologies. If the dictionary is finite, then sooner or later a circularity will occur, and the entire definition will collapse into a tautology, stating of consciousness that it is what it is. In the end, then, our child will have grasped it only in its original and ineffable unity. If, on the other hand, the dictionary is infinite, the process may never come to completion.

This reductio ad absurdum suggests that human sensation cannot be reduced to an unequivocal act of apprehension, of the type $A = B$, but rather unfolds in a relation of partial identities. For each time a thing B is said of a subject A, it is still necessary, for the relation to make sense, that this thing B possess an intelligible existence independent of the subject to which it is attached. Each time a subject A is defined in relation to B, there also exists, according to the principle of differentiation, a link that separates A from B. But what could this link possibly be, in the absence of any prior concept?

As things stand, the psychic machine knows how to produce nothing other than associations. It follows that to dissociate a subject A from its predicate B is nothing other than to affirm the existence of a link between B and a third element C which does not exist between A and B. Each time two things are brought into relation with one another, they thus conceal a third

term. This term, as if by a sleight of hand, is mysteriously smuggled out of formal language. And yet it is essential to human intuition. So essential, in fact, that this third term C must itself become the pretext for a fourth term D, and so on.

At this point, perhaps the time has come to defend myself against an obvious objection. The famous principle of differentiation I have introduced—upon which so many conjuring tricks seem to rely—would it not itself be concealing an infinite regress? For everything I have just described unfolds rather like Aristotle's argument of the third man.

This argument, let us recall, is a critique aimed at Master Plato. It consists in pointing out, not without a certain mischief, that if every object in the real world were to correspond to a Form in the world of Ideas, then these Forms would have to be infinite in number. Indeed, a first Form would be required to explain what a given man of the real world is, then a second Form to explain the relation between that man and the first Form in question, and so on. The mere mention of an infinite cycle, for the Ancient Greeks, likely exerted the same deterrent force as the hypothesis, for contemporary physicists, of the existence of a soul. Wave it under their noses and you will see them immediately retract with the endearing confusion of a lover caught before a damsel.

Now not only do I believe that the self-referentiality of human knowledge is not identical to the argument of the third man, but beyond this—and this is the most important point—I do not see in what way the infinite should be considered a fallacy. On the contrary, I believe that the infinite can be made into a sensible intuition. Such an intuition reveals itself, for instance, when, standing before the sea from the edge of a cliff or the prow of a ship, one envisions a boundless expanse of water, with neither beginning nor end. Inhaling the dusk-tinted sea spray with a pang of nostalgia, one might then say, with Baudelaire[16]:

> The sea is your mirror; you contemplate your soul In the infinite unrolling of its billows; Your mind is an abyss that is no less bitter.
>
> *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*, in *Les Fleurs du mal*.

What to take away from this chapter

The *intuitive judgement* occupies the middle ground between rational judgement of the top-down type and experience-based judgements of the bottom-up type. It consists of seeking contrasts in the entirety of perception in order to identify objects such as: a colour, a pawn, stone, a column, etc.

In contrast to neural networks, intuitive judgement does not possess minimal perceptive units, such as pixels or bits. It does not merely associate these units with one another through the operation of statistical laws; its task is, on the contrary, to construct units of analysis.

To put it differently, the human mind is composed of three cognitive levels: sensation (bottom-up), perception (intuitive judgement), and rationality (top-down). The transition from sensation to perception is made possible by means of two ingredients:

1. *The intuition of Infinity* is the capacity to think about the relationship between a universal concept, such as the concept of a "pawn," and the multiplicity of individual pawns designated by this concept.

2. *The intuition of the Whole*, on the other hand, is the capacity to think non-sequentially, grasping the entire interdependence of concepts and objects.

These intuitions are the analogue, in Kant, of the intuitions of space and time. My tricameral theory of the mind thus represents a profound revision of synthetic *a priori* judgement.

In the next section of this thesis, I will attempt to characterise these two intuitions more precisely and show how they elude our computers.

Intuitive Judgement in Front of the Machine

Extension of a Turing Machine

Chapter 9

The Dilemma of Intuition

Intuition is ambiguous, it harbours within itself an infinite complexity, and reveals many faces one after the other.

Arthur Schopenhauer

Can a machine think? It is time to examine what the conceptual tools adopted in previous parts may teach us concerning the question at the core center of this dissertation. I have endeavoured — with no small diligence — to justify the existence of three strata in human thought: the level of *percepts* (sensation), the level of *intuitive concepts* (perception), and the level of *rational concepts* (abstract reasoning).

Abstract or rational concepts are defined solely through words or numbers. They are pure syntactic signs. They do not always correspond to any tangible or lived reality, yet they may be manipulated by means of logical reasoning. This level of thought most closely resembles operations of computers — and yet, paradoxically, it is the most superficial. When discussing Gödel's incompleteness theorems, I pointed out that, in any axiomatic theory of even moderate complexity, most propositions are syntactically undecidable: their truth or falsity cannot be determined from the axioms alone. TThe rare truths that logic does unveil are but isolated islets — lonely specks scattered across an infinite ocean. In other words, calculative reason is the skin of the fruit, the bark that veils the sap!

The level of percepts, by contrast, belongs to sensitive world. Perception emerges as a whole: a dense neighbourhood of events and spatial forms interwoven. There are no atomic units of perception — no pixels, no instants — that could serve as fixed coordinates. In the absence of such elementary bricks, the mind must isolate contrasts within the perceptual field, following guided by a principle of differentiation and assimilation among its parts. Forms arise — not always consciously. At most they are fleeting sensations,

sometimes intense, but intimately tied to the context of experience, wherein it becomes difficult to distinguish the subject who feels from what is felt. One cannot isolate the percept and claim it exists outside the sensible experience itself. This layer of judgement is what neural networks attempt to emulate. Statistical correlations are drawn so as to give rise to categories within sensory data. Though the results may be impressive, I have endeavoured to show how these algorithms fundamentally diverge from human perception in structuring the data.

The third level, situated between the other two, is that of intuitive concepts. Here, the human mind isolates a perceptual form and grants it an intelligible, conscious, and linguistically expressible reality. In this process emerges a distinction between the concept and the thing it refers to. One may adopt a Platonic standpoint, in which case the concept exists ideally and sensible forms are but its particular manifestations. But one may one might side with the empiricist, for whom the sensible precedes, and concepts merely synthesise it. Either way, the concept is a reality manipulable by thought, but bound to sensible content. It is neither a pure syntactic abstraction (as in Rationality), nor a mere singular sensation (as in Sensation).

I have suggested that the intuitive concepts of my theory occupy, in relation to Kantian thought, a position analogous to synthetic *a priori* judgements such as "The sum of 5 and 7 is 12" or "The sum of the angles of a triangle is 180 degrees." These judgements are grounded neither in pure reason nor in pure experience, but result from the way human beings structure phenomena. Kant's great insight, to phrase it with all elegance it deserves, is that sensible data are not perceived as they exist *in themselves*, but are shaped accordingly to our prior concepts and intuitions. These latter belong to the innate endowment of human being. When we think the universe is a Euclidean space or time is an objective reality, we behave like a overzealous engineer : having imprinted our tools upon all things, we believe Nature fashioned the world with those very same tools.

If the machine does not think as humans do, it is because it does not structure sensible phenomena with the same tools. To perceive this clearly, let us first observe how things appear within human perception, and then compare this to our algorithms. Within our mind — as I believe I have stated more than once — there exists an intimate relation between the whole of space and its parts. Points in space are not raw data; on the contrary, they draw their intelligibility from their relation to neighbouring points. There is thus an entanglement between space as a whole and the figures that unfold within it. This is what led Kant to claim that space is the receptacle of phenomena, that without it no body could be thought.

A similar, though more elusive, phenomenon occurs in our experience of time. Events are not given upon a chronological line extending from past to future. Human brain — which we attribute many a magical trick —

does not merely line up stimuli in a queue. In the very experience of human mind, there are no instants *per se*, but only transitions unfolding across time intervals. This is why subjective time, for us, on the degree of change within our inner lives: when sceneries shift, when thoughts drift, when emotions run high — time seems to pass more quickly than when we remain in perfect stillness.

By contrast, algorithms handle sensible data in a radically different fashion. To begin with, they possess an internal clock — a mechanism that allows them to perform exactly one numerical operation per unit of time. Moreover, the machine cannot conceive of the relationship between the whole and its parts as interdependent. It slices information into atomic units of perception: an image is broken down into pixels, an audio recording is digitised, words are split into letters and read one by one (whereas human grasp the word in its entirety[1]). The machine's ontology is a *flatland* — a world with no relief — where knowledge is acquired by stacking elementary bricks. To put it in a lapidary phrase: the structures of human perception are holistic, while those of the machine are discrete.

In this third part of my thesis, I would like to revisit two questions that I left in the shadows. The first concerns how exactly to define holistic intuition and the intuition of the infinite, which I have claimed to be foundational to human perception. The second concerns what would need to be changed in the machines we currently use in order for them to attain human-like intuition. These two questions converge upon a common objective: to formalise the notion of intuition and give it a mathematically verifiable form. If, indeed, I manage to represent these intuitions by means of a mathematical function, it will be straightforward to see whether this function is or is not computable by a Turing machine. It would suffice to determine whether, when given the correct input on the tape, it can produce the correct output.

This initial formulation of the question, however, is misleading. It presupposes that intuition is representable by a function, and that this function can be expressed within the language of the Turing machine. But the latter, as we have seen, represents information by discretising it. It is far from clear whether, under such conditions, it could ever possess the intuition of an indivisible point. I will now explain this objection more carefully.

When a human being draws a point on a sheet with a pencil, he knows that this rough dot contains within itself an infinity of mathematical points. He knows this because he spontaneously decomposes the whole into parts; indeed, it would be impossible for him not to continue this process of decom-

[1] A study from the uinervsity of Cbirmadge dmonstaretd that it is psoisble to raed a txet in wihch the ltetres are sracbmeld, so lnog as the fsrit and lsat ltteers of ecah wrod rmeian in the rgiht pclae. This dmeonsatrtes that the brain does not read all the ltetres but pecreives the wrod as a wlohe. Proof: admit you had no trouble reading this sntnece. (For a commentary on this paragraph, see Thierry Opillard, *La Dégradation Cambridge: Même pas vrai?*, Les Actes de Lecture no.102, June 2008.)

position, at least in imagination. The computer, by contrast, begins from the principle that there exists an indivisible point: the pixel.

This quest to define intuition inevitably leads us to the question of the continuum: can a computer think a shape, a segment, a line, in its entirety? Can it, like the human being, represent the relation that binds a form as a whole to each of its parts? Or can it only represent space as the agglomeration of indivisible pixels?

These preliminary questions lead me to a first hypothesis, which I shall attempt to put to the test throughout the following pages:

Hypothèse 4. *Only a Turing machine capable of thinking the continuum, or of thinking the whole, could have access to the holistic intuition and the intuition of the infinite that characterise human perception.*

9.1 The Intuition of Infinity and the Intuition of the Whole

The Turing machine is the paradigmatic model of the computer, which I have described in detail in section 5.3.1. It consists of an infinitely long tape, on which a reading head inscribes sequences of symbols such as 1 and 0. This head—nothing less than the sacred stylus of computer science—can read existing symbols, erase them, write new ones, and move freely along the tape. The question I raise here is whether this Turing model could be capable of constructing intuitive concepts. What has been laid out in the preceding section of this thesis allows me to identify at least four theoretical obstacles:

- **Discretisation:** Within a Turing machine, time is composed of a succession of discrete moments, and the tape is made of a succession of cells. Thus, at each unit of time, the machine can carry out exactly one operation among the following: move from one cell to another, read a symbol, or change a symbol. Yet in the human mind, it would appear that points in space do not exist prior to the act of perception. I have demonstrated, through various illusions, that the mind apprehends the whole not by scanning it centimetre by centimetre, but by identifying the contrasts between this whole and some of its parts. When confronted with a photograph, it does not begin by sweeping through the pixels from left to right and top to bottom. Its gaze follows a trajectory that is indeed predictable and orientable, but which depends on the contrasts and saliencies within the image. Points are not independent individual realities; they derive their intelligibility from the whole to which they belong.

- **Sequentiality:** The Turing machine possesses an internal clock. It performs operations sequentially, moving from a point A to a point B. Yet I have sought to argue that time, as experienced by human beings, is not ordered in an absolute manner. Experiments on chronostasis show that when two stimuli are separated by less than 100 milliseconds, we may sometimes invert their order. This phenomenon is not a conscious decision of the will. It is carried out automatically when certain conditions are met. Everything suggests that human perception does not attend to sharply defined instants in succession, but rather to neighbourhoods of temporal events. It is more sensitive to the contrasts between phenomena within that neighbourhood than to the absolute durations separating them.

- **Univocity:** The Turing machine is defined by univocal assignments. When one writes $A = a$, it means that the value a is assigned to the variable A, or in other words, that the Turing machine has inscribed the symbol a in cell A on the tape. This notion seems to be constitutive of computability. After all, is a computation not precisely the act of finding the value of a certain variable?

  Yet in the human mind, it is not at all clear that the relation of predication is univocal. The proposition "this is a tree" teaches us something meaningful about the nature of the "this" that stands before us—but also about what it is to be a "tree". The subject of the proposition (this) and its property (being a tree) are commutative.

  In other words, it is not by reading the dictionary definition of a tree that we form an accurate idea of what it is. It is by walking through the forest, comparing trunks, leaves, branches, by exposing ourselves to the full sensory multiplicity contained within the concept. This implies that when we think of the concept "tree", we may represent a series of distinct trees. These trees share similarities, but none are perfectly identical. One could thus think that the human being, even when constructing a concept as seemingly banal as "tree", is manipulating infinite possibilities. It synthesises this infinity into a single mental image.

- **Absence of Semantics:** This human capacity to grasp the infinite is especially evident when we represent to ourselves the set $\mathbb{N}$ of all natural numbers. We cannot explicitly enumerate the integers $1, 2, 3, \ldots$—our memory is far too limited to do so. Yet there exists in our mind a truly extraordinary faculty that allows us to generalise the number sequence and asymptotically approach ever-larger magnitudes. This takes place through a process of analogy.

> Suppose we wish to form an idea of the number one thousand, or one hundred and fifty million, or a one followed by seventeen zeros, in order to approach the idea of infinity. To do so, we might begin by imagining a one-centimetre ball placed in the centre of a football pitch. We then realise that one thousand is equivalent to the ratio between the diameter of this ball and the size of the pitch. We go on to imagine that if this ball were the sun, the 150 million kilometres separating Earth from the Sun would be comparable to the distance from this ball to the edge of the pitch. Then we might think that the 28,000 light-years (i.e. 10^{17} kilometres) separating the Sun from the centre of the Milky Way are comparable to the distance from our ball to the Sun. By repeating this mental experiment step by step, representing ever greater distances, we eventually reach the dizzying idea of the infinite.
>
> The faculty to imagine an unrepresentable idea—now that is something that seems to elude the Turing machine! It may indeed possess an infinite tape on which to inscribe the numbers $1, 2, 3, \ldots$, but can it form an idea of what lies at the end of this sequence? When it writes the symbol $\mathbb{N}$, enumerates its properties, and manipulates it within a proof, it is confined to a rational, abstract, purely linguistic definition. It understands syntax—but what of semantics? Does it not behave like a monkey handling words without understanding what they denote? Granted, it is a precise and exact monkey, a learned monkey—but is this mind endowed with human *understanding*?

In the pages that follow, I will propose that there are two forms of intuition: holistic intuition, concerning the relation between whole and parts, and transcendent intuition, relating to infinity. These dimensions of intuition are the anvil upon which the human mind, wielding the hammer of experience, can forge its categories. Without this anvil, experience would be but empty motion. It needs resistance in order to give rise to objects. Intuition, and its many facets, is thus the faculty by which objects are apprehended—not by means of a traceable, logically deductive reasoning that proceeds linearly from premises to conclusions, but in a global manner. If we trust its Latin etymology, *intuition* comes from *intuitio*, meaning "the act of seeing in a single glance." There is indeed a tendency in philosophy to oppose the concept of intuition to that of reason. One need only consider this sentence by Schopenhauer: "Concepts form a special class of representations, entirely distinct from intuitive representations such as those we have discussed thus far, for they exist only in the human mind[2]."

The first form of intuition, *holistic intuition*, also known as immanent

[2]Schopenhauer A., *The World as Will and Representation*, Chapter 9

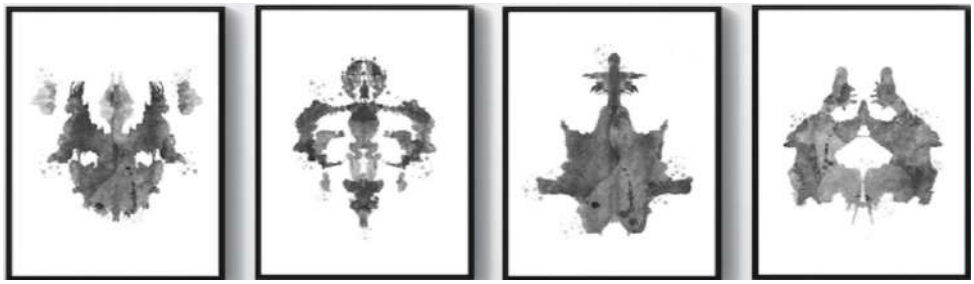


Figure 9.1

intuition, may be described as a certain faculty for grasping the impermanent relations between concepts. Whereas the rational mind lays words out flat, attempting to construct hierarchical relations among them, immanent intuition operates cyclically. Reason simplifies the problem, fixes its initial elements, and deduces conclusions in the Cartesian manner of the *Discourse on Method*; intuition complicates, weaves relationships, and opens perspectives more than it offers answers. In psychology, this immanent intuition is sometimes referred to as "branching thought," "divergent thinking," or "right-brain thinking." These terms, however, carry variable and somewhat ambiguous definitions, and—with the exception of the second—are not firmly established in the scientific literature. Someone endowed with this form of intuition in an acute way will tend to distrust concepts, finding them inadequate to grasp the richness and density of relations between things. The very idea of decomposing the world into objects with individual existence, rather than representing it as a whole, will seem questionable.

This holistic intuition is the principal ingredient of Sensation, when it produces percepts. These do not exist independently. They are not stable individual objects, identifiable outside their context, but contrasts that emerge within a whole. This is why, when you are shown coffee stains on white paper—resembling a Rorschach test—forms arise in your mind (see figure 9.1). You may see a face, a monster, some characters, mountains. These contours exist only indeterminately. If you change the background or rotate the paper, they may transform. It is due to this holistic intuition that the structure of human perception, unlike that of the Turing machine, does not seem to follow a discrete and sequential representation of the world.

The second form of intuition, *the intuition of infinity*, also called transcendent intuition, is the faculty to represent the unrepresentable—to understand that symbols denote something beyond what they contain. In contrast to rationality, which manipulates the signs of reasoning and arranges the words of a sentence as a gramophone arranges phonemes, transcendent intuition grasps that the mental image associated with the concept "tree" is not the concept itself. This is what enables it to introduce a distance between words and things, between signifiers and signifieds. This aptitude

becomes particularly manifest in the capacity to grasp ideas—such as the idea of infinity. It is indeed curious that such an idea appears in us from early childhood, whether we are thinking of the external world (where bodies extend through an infinitely extensible space) or of our inner consciousness (where phenomena unfold in an infinitely extensible temporality).

Transcendent intuition, logically enough, is already present in embryonic form at the level of percepts. For these percepts are composed of a potentially infinite number of parts. The Rorschach blots can be divided into sub-blots; one can see shapes within the initial shapes. Yet infinity truly emerges only when this intuition manages to unify these categories under a single image by saying of such forms: “It is a man!” or “It is a tree!” It is then that the idea arises that the concept “tree,” as an intuitive concept, overflows all known and imaginable perceptions of trees. It is thanks to transcendent intuition that the human mind can give meaning—a semantics—to the symbols it manipulates[3].

9.2 Alan Turing’s Objection

It is one thing to declare that a Turing machine lacks intuition, but quite another to demonstrate it. The moment intuition is defined in contrast with logical reasoning, it becomes elusive to articulate its essence rationally. If we fail to specify, rigorously, what this intuition entails and merely proclaim vaguely, for instance, "A machine cannot possess genius," we are led back to questioning the essence of genius itself. Might it be a catch-all concept, much like idealists see in the soul the ultimate stand against materialism? In his renowned article, *Computing Machinery and Intelligence*, Alan Turing had already foreseen this issue[4]:

> *"These arguments take the form, "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X." Numerous features X are suggested in this connexion I offer a selection: Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have*

[3] This is a purely technical remark, but it is worth noting that one can circumvent Gödel’s incompleteness theorems by appealing to a transfinite proof. It is as though the human being can give meaning to certain unprovable propositions because its semantics—or its experience of the world—is not bounded by a finitist representation. This point would, however, require more extensive development. I shall return to it in more detail when I define transfinite Turing machines.

[4] Translation is my own. The passage is from the section "Arguments from Various Disabilities."

as much diversity of behaviour as a man, do something really new."

Among these X-listed qualities, many seem vague or hard to define. Is it not this conceptual haze that tempts us to think that novelty, for example, resists algorithms? If a mathematical criterion were provided to determine if a particular artistic creation (a drawing, a poem, a piece of music...) statistically diverges from a set of known productions, then no reason would exist to doubt that a neural network could be trained to produce novelty. It might even be that the machine's novelty would soon surpass that of a human.

Yet, I can already hear my detractors philosophically mumbling: "Novelty is not merely statistical independence, but requires another essential ingredient." What is this ingredient, then? Genius? Originality? A paradigm shift? In our quest to explain the novelty's X in verifiable terms, we stumbled upon a new and mysterious X.

I dub this issue the *dilemma of explicitation*. It unfolds as follows: if we assume a certain faculty (such as intuition) distinguishes the human mind from a machine, then

- Either this faculty is defined in explicit, verifiable terms, in which case nothing prevents us from believing a machine may mimic it;
- Or this faculty isn't defined explicitly or verifiably, and we may doubt its existence.

Note, this dilemma extends beyond intuition to any faculty invoked to draw an inviolable line between humans and machines—be it memory, emotion, subjectivity, etc. Intuition is distinct in that it often signifies a knowing ability that opposes reason's concepts. It stands as a favoured point of resistance to the idea that the mind, by nature, is nothing but a set of computable operations. Schopenhauer notably illustrates this opinion:

The essence of true wisdom, each person's true science, is not confined within concepts, in abstract knowledge, but is entirely in intuition and in the degree of penetration, accuracy, and depth with which one has grasped this intuition.

Consider also that certain definitions of the human mind are explicit, yet unverifiable. It is asserted, for example, that only humans can feel sensations or possess an inner world. Deciding this interiority's existence in a living or artificial being seems to resist an obvious answer, as people often hold markedly different views on such matters. Thus, I shall first address intuition in verifiable terms, questioning, by the end of this part, if the conclusions drawn can illuminate more fundamental metaphysical problems.

Unsurprisingly, in light of this dilemma, there is a polarisation in the debate on artificial intelligence: optimists argue that many tasks, once deemed beyond machines' reach, have been accomplished—and those remaining will likely follow suit. Pessimists, however, develop definitions that are not mathematical enough to test against machines.

To transcend this dilemma, philosophy is called to open a third path. This would involve defining intuition in a way that, while perfectly explicit, evades calculation. Although this idea may seem puzzling, it can indeed be likened to a mathematical analogy: in analysis, differential equations exist without analytical solutions. One can show such equations have exact solutions, yet synthesis into a computable formula eludes us; thus, a computer can at best offer approximations.

So frequent are these equations in science, that exact solutions are exceptions. In classical physics, for instance, gravitational effects on systems comprising more than two bodies cannot be calculated perfectly; and in quantum physics, Schrödinger's equations only perfectly solve for the hydrogen atom. Assuming the metaphor applies to the human mind, and that it surpasses the complexity of a hydrogen atom, might some integrals in mental processes, while perfectly defined, defy calculation?

The solution I propose in this thesis's third part will design a mathematical formalism to express the intuition of infinity. Such formalism should be as operational as possible, yet elude a classical Turing machine's functioning. I'll demonstrate its potential for dialogue with neuroscience, particularly regarding structural analogies between brain function and phenomenological experience. Before embarking on this challenging task, I shall first show that certain definitions of intuition documented in literature, especially those crafted in response to artificial intelligence challenges, falter upon the dilemma of explicitation.

In this chapter, I shall discuss Hubert Dreyfus's definition, which significantly impacted research paradigms in *intelligence learning*. I'll reveal how this definition makes an intriguing distinction between knowing (know-that) and doing (know-how). I'll note, however, that since algorithms embraced neural networks, evolutionary programming, and other machine learning methods, it's clear machines can implement this definition of intuition.

I will then consider two classical strategies from literature to navigate out of the dilemma of explicitation. The first, termed *mysterianism*, posits that intuition exists but is undefinable by human intelligence. This is a denial of the explicitation dilemma's relevance and needs careful study; for, if valid, any attempt to define intuition would fail, and might partially invalidate results achieved via my mereological theory of consciousness. When McGinn popularised this philosophical stance, it concerned not intuition, but consciousness itself.

The second strategy, much older, supposes the existence of intellectual intuition. Traceable to Plato, it has served philosophers as a straw man

opposed by almost all modern traditions. More precisely, after Kant decided to eliminate it, he dissuaded anyone after him from resurrecting this concept. Yet it could offer a genuine alternative to the explicitation dilemma, having the characteristic of being definable in philosophical language but not in logical-deductive terms. Plato argues it allows contemplation of forms or ideas beyond *dianoia*, which is but *analytical intelligence*. It presents an ideal candidate to challenge the algorithmic ambition of reducing the field of intelligible truth to the realm of calculation.

Following this journey through existing philosophical literature, I shall attempt a personal synthesis. I will, indeed, concede that the intuition of infinity, crucial as I have shown it to be for consciousness's function, cannot be expressed in Turing Machines' mathematical language but can nevertheless be formalised in higher-order machines. These machines, which I term transfinite Turing machines, share the intriguing capability with Chuck Norris of counting to infinity... multiple times. I will provide a concrete example of an NP-complete problem where both humans and transfinite Turing machines can think holistically, while a classical Turing machine remains tethered to term-by-term analysis. I propose this may be explained by the intuition of infinity, whose definition I have introduced.

9.3 Intuition according to Hubert Dreyfus

The definition of intuition provided by Dreyfus appears within a very particular intellectual context. In 1958, the effervescent climate of American research, marked by the face-off between two nuclear powers, was favourable to the development of information technologies, including artificial intelligence. Herbert Simon had indeed written that, within less than a decade, the computer would (1) be able to establish its throne in the realm of mathematical ideas by proving new theorems, and (2) challenge its creator's face by defeating human champions on a chessboard.

Ten years after this bold prophecy, alas, the scientific community was forced to admit that computer-based mathematical proofs were increasingly resembling facsimiles rather than the brilliant genius of Carl Gauss[5]. As for the victories on the chessboard, they were largely due to the speed of the computer's calculations at the end of the game and did not hold up much outside of blitz games. It was therefore tempting to think that master-

[5]In 1978, exactly ten years after Herbert Simon made his "prophecy", Haken, Appel, and Koch published an article demonstrating that an IBM program, of their design, could verify the four-colour theorem across a billion instances. Although the verification of the program was controversial, and many operations in this "proof" remained in human hands, they were attempting to suggest the theoretical possibility of an automatic proof assistant. It would take until 1996 for the scientific community to reach a first consensus, when the proof was reduced, by Robertson, Saunders, Seymour, and Thomas, to the analysis of 633 combinations sufficient to provide a conclusive proof.

ing a complex cognitive task, far from being reduced, like students reciting their multiplication tables, to pure symbolic calculation, involved "intuitive" thought processes that eluded classical computational paradigms.

It was with this presumption that Herbert Simon's adversary, Professor Hubert Dreyfus, sought to outline a theory of learning based on the distinction between declarative knowledge (know-that), easily manipulable by machines, and procedural knowledge (know-how) unique to humans[6]. In this theory, presented in 1986 in *Mind Over Machine*, cognitive evolution within a discipline progresses through five distinct stages, from the simple understanding of explicit rules to the implicit, and in a sense *intuitive*, mastery of complex configurations[7].

I will briefly outline these five stages, for not only do they represent a fine definition of what human intuition *could* be; but they have also significantly contributed to the development of artificial intelligence. The birth of neural networks, which we will touch on in later chapters, can be seen as the quintessential historical attempt to implement, within a new computational paradigm, the learning ideal depicted by Dreyfus.

- **Novice Stage**: At this first stage, the subject acquires a certain number of factual pieces of information about the discipline, but identifies objects and actions to be performed without reference to context. For example, a chess player who learns how to move the pieces and uses the rule of thumb. According to this rule, a queen is worth twice as much as a rook, which in turn is worth five times more than a pawn — regardless of the relative positions of the rook or pawn.

- **Advanced Beginner Stage**: The subject can identify relevant situations or properties without being able to describe the rules by which this is so. He uses this intuition to complement the rules acquired earlier. A chess player at this learning stage would recognise an advantageous pawn structure, or the presence of weaknesses in the king's position, based on accumulated experience from previous situations.

[6] It should be noted that Dreyfus does not deny the possibility of a machine acquiring procedural knowledge (know-how). He merely states that, given the state of knowledge in his time, it cannot yet be imitated by computers. It was precisely due to Dreyfus' critique that programming methods would evolve, ultimately leading to machine learning and neural networks.

[7] This theory, Dreyfus specifies, was empirically verified by Patricia Benner. She collected data on nurses with varying levels of experience (Patricia Benner, *From Novice to Expert: Excellence and Power in Clinical Nursing Practices*, Addison-Wesley, 1984). This model has so profoundly influenced later development that it is now widely integrated or debated (see, for example, Peña Adolfo, *The Dreyfus model of clinical problem-solving skills acquisition: a critical perspective.*

Nevertheless, they continue to follow explicit, hierarchical rules.

- **Competent Stage**: The subject is capable, when facing any situation, of redefining the hierarchy of operations to be carried out in view of a higher goal. They are no longer concerned with the rules themselves but act according to the ad hoc plan they have set for themselves. A chess player at this stage sees the board as a set of tensions leading to positional goals. For instance, when they believe it is time to launch an attack on the king's side, they will concede certain weaknesses in their defence. They act without consciously mobilising the rules for moving pieces or the rule of thumb.

- **Proficient Stage**[8]: The subject acquires automatisms. It is no longer necessary for them to rationally think through these automatisms. However, they must compare the different automatic responses considered, or engage in analysis to define how to conduct their overall strategy. A chess player, relying on their knowledge of numerous analogous situations, will instinctively have a list of possible moves in a given situation, allowing them to play very quickly without a significant drop in performance. Dreyfus adds that players who see chess as an analytical game typically remain stuck at stage 3 of the learning process. He cites the testimony of one of his colleagues[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 psycholigist 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].

- **Expert Stage**: The expert has such a fine knowledge of situations that they respond automatically. They are no longer truly aware of manipulating pieces on a chessboard, but are immersed in the task. They feel the stakes far more than they objectively think through their actions and decisions. If they begin to think about the colour

[8] In English, Dreyfus uses the term «Proficient», which doesn't have a widely agreed-upon French translation. I decided to replace it with the term Performant to capture the idea of advanced and beneficial skill. Its Latin etymology stems from the word proficiens, present participle of *proficere* ("to advance, make progress, succeed, be profitable or useful").

[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.

of the pieces or the fact that they are playing, it takes them out of their state of mastery. "When things are proceeding normally," Dreyfus specifies, "experts do not solve problems or make decisions; they do what works ordinarily[11]." As an example, he gives that of Kaplan, an international grandmaster who was always able to play chess correctly at five seconds per move, even though he simultaneously had to perform arithmetic operations. This grandmaster recognises game configurations like anyone would recognise, if exposed to it, the face of a close person. According to Dreyfus, this automatic holistic understanding is made possible by a cognitive approach that does not seek to reduce the whole to its parts. In this sense, he practices a form of arationality[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.*

It should be noted that the development stages above do not necessarily align with the individual's level of performance, but rather with types of cognitive functioning: in principle, a talented novice can defeat a less inspired competent player. This is perhaps why a chess prodigy like Paul Murphy could, at a young age, learn the rules by watching his father play, but also explain to him at the end of the game that he had made tactical mistakes. In certain contexts, the perspective of the novice may present advantages over that of a more experienced individual, especially when the former possesses an extraordinary memory, dexterity, or fluid intelligence. Far from seeing these five stages as levels of competence, they should be viewed as modes of cognitive functioning, which Dreyfus assumes to follow one another as the human subject gains experience.

The fact that Dreyfus illustrates these five stages through the game of chess is undoubtedly a deliberate mockery aimed at the overly optimistic prophecies promoted by the artificial intelligence establishment[13]. For the record, Dreyfus was defeated on the chessboard, ten years before the publication of his book, by Richard D. Greenblatt's Mac Hack VI software.

[11] *Ibidem*, p.31.

[12] *Ibidem*, p.36.

[13] It is worth noting that Dreyfus provides several other examples, notably that of driving. To illustrate the novice level, he also speaks of letter recognition: an individual who recognises the letter "E" by identifying the three horizontal bars behaves like a novice. Someone who recognises it by habit, or more precisely, by *pattern recognition*, is an expert (*Mind Over Machine*, p.22). Ironically, these two examples constitute symbolic historical steps in the more recent development of neural networks.

This debacle, which the famous Herbert Simon witnessed, earned Simon's enthusiasm: "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 games[14]." Nevertheless, this did not prevent the defeated philosopher from seeing in his colleagues' software nothing less than a "novice" with great computational ability. A sort of Paul Murphy of cables and silicon.

The idea that human intuition continues to elude thinking machines, for Hubert Dreyfus, has its roots in the Heideggerian conception according to which the human being is a "being-in-the-world". The essential property of such a being is to sense the world within itself, to the point that one could say the world is born and dies with it. Now, since it is the fate of every human being to disappear and become aware of the contingency of their existence, the question they face is the following: why does the world, this world they carry within them, exist, when it might not be? The human condition is thus permeated with a metaphysical question — a question without an answer — which drives the human being to weave relational answers with their environment. Each object not only has a logical meaning, but is also an opportunity for the subject to affirm their unique existence in the world. The cane, in the hands of the blind person, is immediately integrated as an extension of their senses, whereas for a man with impeccable sight, it will rather serve as an aide to the action of the motor muscles. It is also bearer of memories, automatic associations of ideas, habits, doubts, and hopes. The tool, in the hands of the Homo faber, is not off-the-shelf. It is the custodian of an implicit know-how, which in psychology is called affordance, and which Dreyfus sees as the foundation of intuition.

This claim that know-how is, in the current state of knowledge, inimitable by a machine; and that this makes computers unsuitable for performing certain tasks in which, however, humans can excel, will unfortunately suffer a major setback. When Deep Blue, after several crushing defeats, finally wins the match against Kasparov, it must be admitted that a computer can reach, or at least imitate, the highly praised expertise.

Certainly, one might doubt that the machine, even in defeating the human at chess, acts with the same kind of automatism. Isn't the computer's automatism, far from consisting of an accumulation of concrete experiences, simply a set of dumb and brutish calculations? Yet, one thing remains indisputable: the know-how introduced by Dreyfus inspired a new generation of computer scientists. Artificial intelligence programs, abandoning the plough of symbolic calculation, have seized the steam engine of intelligence learning. Taking advantage of what software could do with a vast database to correct its errors, to become sensitive to context, to spot patterns, and to increase

[14] *Oral History of Richard Greenblatt*, interviewed by Gardner Hendrie, 2005 Computer History Museum, CHM Reference number: X3131.2005.

its capacity for interaction with the environment, the computer suddenly became eligible for the very intuition that had been denied it all along. We thus moved from the symbolic paradigm, based on logical calculation, to the connectionist paradigm, based on the manipulation of large data libraries. In the face of this transition, the philosophy of science is put to the test, either to redefine what intuition is, and to explain again how this intuition distinguishes the human being from the computer, or to renounce seeing, in this sacred intuition, a dividing line, or at least a point of friction, between human faculties and artificial faculties.

9.3.1 Objections to Dreyfus's Position

We thus almost incidentally come to the crucial question: assuming that the study of human intuition offers some insight into epistemology, is it possible to describe what it is without this description, once formulated, being, at least theoretically, simulable (or emulable) by a computer? When Dreyfus defended the existence of know-how, he argued that the activity of the mind was not merely reducible to a set of algorithmic operations, but also relied on the recognition of implicit patterns. The difficulty of this position, however, was in describing what know-how was in sufficiently explicit terms to convince us of its existence, without simultaneously reducing our knowledge of this know-how to a mere know-that that could be imitated by a calculating machine. This problem is what I have called the *dilemma of explicitation*.

Everything ultimately happens as if the recent history of the philosophy of artificial intelligence, regarding intuition, had this paradoxical double effect: on the one hand, it confirms the proponents of strong artificial intelligence in their belief that it is ultimately possible to imitate, through a mechanism, all the functions of thought traditionally associated with the concept of human intuition; on the other hand, it also confirms the opponents of this view in their belief that intuition cannot be described in *purely* explicit terms (although descriptions may shed light on it).

Let us clarify this idea. The proponents of strong artificial intelligence are confirmed in their optimism for this simple historical reason that, every time a definition of intuition is presented, and we can think, for example, of sensitivity to context, this definition ends up succumbing to the machine's insatiable progress. Thus, the driving of an automatic vehicle in a real urban environment, or the fine understanding of a sentence's semantics, once considered unattainable with a simple machine, have today become so commonplace that no one would now deny their automation. A sceptical reader need only ask ChatGPT[15].

[15] I indeed addressed this question to ChatGPT 4.0 in November 2023, and here is the result of the conversation: "Question: Is the fine understanding of the semantics of a sentence an unattainable task for machines?"

Answer: The fine understanding of the semantics of a sentence is a complex task that

At the same time, and as oxymoronic as it may sound, the opponents of artificial intelligence's technical promises are further bolstered in their scepticism. Indeed, the repeated failures to provide a definition of intuition that resists mechanisation, far from being proof that humans function like machines, are, in their view, so many attestations to the impossibility of formulating, in mathematical terms, a mode of thought that lies beyond apodictic reasoning. Is it not the essence of intuition to escape logic? The heart of thought, so to speak, has its reasons that reason knows not. Pascal also says about the truths of the mind[16]:

> *We must see the matter at once, at one glance, and not by a process of reasoning, at least to a certain degree. And thus it is rare that mathematicians are intuitive, and that men of intuition are mathematicians, because mathematicians wish to treat matters of intuition mathematically, and make themselves ridiculous, wishing to begin with definitions and then with axioms, which is not the way to proceed in this kind of reasoning.*

9.3.2 Challenging the status quo

A simple and elegant way to resolve this fundamental contradiction, without falling from Charybdis to Scylla, would be to define intuition within a framework that, although perfectly explicit, would be beyond the reach of any possible machine. Such a strategy can be articulated in two obvious ways:

1. The concept of intuition cannot be explained through language;
2. The concept of intuition can only be explained through a non-mathematical language;
3. The concept of intuition can be explained through a mathematical language, without being achievable by a machine;

---

has long been considered unattainable for machines. However, recent advances in the field of machine learning, particularly with deep learning-based natural language processing models like GPT-3, have significantly improved machine performance in this task.

These models can analyze vast amounts of textual data and learn sophisticated semantic representations that allow them to understand the meaning of sentences with high accuracy. Tasks like detecting irony, sentiment analysis, or understanding nuances in language can now be accomplished with precision comparable, or even superior, to humans.

However, it is important to note that fine semantic understanding remains a challenge for machines in certain situations, particularly with ambiguous contexts, complex sentences, or languages with very different syntactic structures. In these cases, machines may struggle to extract the precise meaning of the sentence and may produce erroneous results. Nevertheless, recent progress is promising, and machines' fine semantic understanding continues to improve. »

[16] Pascal Blaise, *Pensées*, 305.

The first stance is that of mysterianism; the second is that of Platonism. These will be the subjects of the following two sections. The third stance, finally, is the response I will attempt to construct after briefly outlining the difficulties of the two previous approaches. It will present itself as the middle way. It is the synthesizing element that skilfully brings together thesis and antithesis.

Before going further, I must make an observation. Just like the sweeper who discreetly slides the bothersome dust under the carpet, the dilemma, as I have written it, seems to conveniently overlook the idea of simply abandoning the concept of intuition, and affirming — why not? — that the mind is nothing more than a sophisticated computer. I believe that, in the contemporary doxa, this is not the least common opinion. However, insofar as the role of this work is to build an argument to define, if they exist, the intrinsic theoretical limitations of the artificial intelligence programs, the dilemma we have exposed seems unavoidable. Only by seriously considering the idea that human intuition can be explicated will we see whether or not this intuition resists Turing machines.

One might also argue, in a similar manner, that it is not intuition that plays a key role in the face of the computational power of machines, but rather memory, corporeality, or intentionality. By proceeding in this way, one still does not escape the central problem. Whatever faculty X is exhibited to block mechanisation, or whatever name one wishes to give this faculty, it will have no choice but to be either defined in calculable terms and actually executable using an algorithm, or in calculable but non-executable terms, or in non-calculable terms, or not defined at all — or at least not explicitly.

9.4 First Solution: Mysterianism

In the previous chapter, I spoke of the dilemma of intuition while concealing this response under the carpet: that intuition cannot, conceptually speaking, be defined in objective and scientific terms. This alternative is appealing, especially if we accept that intuition has something unrepresentable, closely linked to human consciousness. This connexion between consciousness and intuition is motivated by the fact that intuition imposes itself upon us as something we are conscious of, but cannot always explain. The lack of explanation is partly due to the fact that we cannot trace the unconscious processes that gave rise to this intuition. It is also due to the fact that we cannot explain consciousness itself through clearly defined physiological mechanisms.

I do not wish to give the impression here of replacing one mystery with another, even more obscure than the previous one. I merely wish to say that there are mysterious things within the human mind. The question is

not so much whether intuition is comparable to consciousness, but whether intuition might be one of those mysteries that we experience but cannot explain. Alan Turing himself wrote: "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"[17]. Don't we similarly experience difficulty in localising intuition within the human brain?

In this section, I would like to discuss the *mysterianist* movement, which defends the view that the problem of consciousness cannot be solved by the human mind. To do this, I will begin by tracing the origins of mysterianism as a philosophical movement, starting with the inaugural speech of Emil-Dubois Reymond. I will then attempt to outline a more contemporary position.

9.4.1 Ignoramus et ignorabimus

In 1880, in a speech to the Royal Academy of Sciences of Prussia, Emil du Bois-Reymond listed seven problems which, according to him, could never receive a scientific explanation. Among these immortal monsters of human thought are the nature of matter and force, the origin of motion, the origin of life, the apparent teleology (or finality) of nature, the origin of simple sensations, free will, and the origin of intelligence itself. With this inaugural gesture, Du Bois-Reymond introduced *ignoramus et ignorabimus*: we do not know and we will never know.

It must be acknowledged that the history of ideas seems to favour Du Bois-Reymond. Is it not true that certain fundamental problems, whose list could be discussed, have not seen substantial progress since their formulation in the limbo of Greek philosophy? That some questions have the power to plunge the mind into metaphysical torpor? However, it remains curious to note that, even if the answer to these questions does not exist, we are still able to formulate them. They are therefore not unthinkable. One is thus entitled to ask how it is that a question, for which there is no answer, can still be posed? Would it not be more logical to think that if the question were truly impossible, it would be equally impossible for us to conceive it?

Du Bois-Reymond behaves like those athletes who take a run-up before stopping at the precipice. To push his approach to the limit, wouldn't he have had to ask, in a way that is both more modest and more radical: can we even know whether there is or isn't an answer to all problems? And can we know if there is an answer to the question of whether there is an answer? Still, the state of perplexity was so evident, fifty years after Du

[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.", in Turing A., *Computing Machinery and Intelligence*, p.12. The translation is mine.

Bois-Reymond's speech, that the mathematician Hilbert took the initiative to cut the Gordian knot: "There is no ignorabimus in the natural sciences. We must know, we will know."

It is important to distinguish mysterianism from two other related positions that are not *stricto sensu* equivalent: Aporia and Mysticism. The commonality between these two attitudes is that they recognise there are questions without answers. They thus admit the remarkable fact that we can formulate a question about something we do not know; however, they explain this fact in very different ways:

1. Aporia: The questions we ask are formulated in contradictory terms. If they admit no answer, it is because they do not constitute a true problem. However, we are not intellectually agile enough to spot the contradiction hidden behind all our futile questions. If our language were purified, there would be no false problems.

2. Mysticism: The questions we ask, although perfectly coherent, sometimes admit no scientific answer, no philosophical answer, or indeed no answer at all. The reason for this is that reality is not entirely intelligible. We can only apprehend the mysteries of the Universe through non-rational faculties.

3. **Mysterianism:** The questions we ask do admit answers, but our human intellect is not in a position to understand them. Mysteries thus appear as a highlighting of the intrinsic limits of our language.

Mysterianism, therefore, does not merely state that there are questions without answers. It makes the bet that answers, although they exist, are beyond the reach of our mind. To illustrate this position, and connect it to our dilemma of intuition, I would like to examine a text by the British philosopher Colin McGinn, titled *Can We Solve the Mind-Body Problem?* Although this text concerns consciousness, not intuition, I will focus primarily on the argumentative structure.

9.4.2 McGinn's Stance

In an article titled *Can We Solve the Mind-Body Problem?*, written in 1989, Colin McGinn quotes Julian Sorell Huxley with a certain eloquence[18]: "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 Djin, where Aladdin rubbed his lamp in the story.". With these words, he challenges an entire tradition of explanations of consciousness.

[18] Julian Sorell Huxley (1887–1975), evolutionary biologist and eugenicist, should not be confused with Aldous Huxley (1894–1963), the author of *Brave New World*.

To ground his views, he begins by stating that there are two possible orders of explanation. The first, which he calls the *constructive* approach, involves searching for a set of properties that are satisfied simultaneously in both mental states and in the brain (or the body). The second approach, referred to as the *dualistic* one, and historically dominant in his view, postulates the existence of supernatural entities or divine intervention. It thus appears that the human mind is placed between two solutions, both as mysterious as the phenomenon to be explained: a Cartesianism of substance or a pre-established harmony in the Leibnizian sense.

He then draws the reader's attention to the possibility of a cognitive closure within the human mind. This would mean that there are properties which, although scientifically accurate and explainable, cannot be comprehended by our species. Cognitive closure, in strict terms, should not be confused with perceptual closure, which refers to the ability to apprehend a phenomenon through our senses. For example, there is no doubt that our perceptions are closed off to the atom, which cannot be observed through a microscope. This does not, however, mean that the atom or invisible frequencies cannot be conceived in the human mind. On the contrary, the atom, in its primary meaning – that is, the elementary explanatory unit – seems so essential to the construction of human theories that it did not require contemporary physics for us to be exposed to its concept.

It is worth noting, McGinn adds, that for a classical Humean mind, ideas should form from impressions, so that everything closed perceptually should also be closed cognitively. A similar idea is attributed to Locke, who "thinks that our ideas of matter are closely constrained by our perceptions, and thus concludes that the true science of matter is eternally beyond us – and that we will never be rid of our perplexity regarding the question of what solidity is[19]". But although the human mind is probably not Humean in this restrictive sense, it must still conform to certain principles binding impressions to ideas, and ideas to one another, so that, since the mind is not able to apprehend all sensory phenomena, it is reasonable to think that it cannot conceive all possible ideas either.

The question that then arises is as follows: is consciousness cognitively closed for the human species? Such a position, in McGinn's view, does not imply mysticism. The phenomenon of consciousness itself is not beyond the scope of science – in fact, the philosopher seems to insist on this point with considerable vehemence. In the end, consciousness is like life. Unless one assumes the intervention of God's finger, it is reasonable to think that life formed on Earth from known organic matter. Nevertheless, it could be that the explanation is beyond the reach of the human mind, just as the theory

[19]This quote is my translation. The original sentence can be found in McGinn Colin, *Can We Solve the Mind-Body Problem*, 1989. In a footnote, McGinn clarifies that this idea of Locke comes from his *Essay Concerning Human Understanding*, Book II, Chapter IV.

of relativity is beyond the comprehension of a five-year-old child.

Of course, an argument in favour of this thesis would be to highlight the failures that science has experienced on this difficult question. Such a strategy is analogous to the pessimistic meta-inductive argument sometimes used against metaphysical realism, which posits that there is a world beyond our senses.[20]. However, inasmuch as many scientific phenomena once thought insoluble were ultimately understood by the human mind, this is by no means a significant argument.

McGinn then prefers to opt for an approach that bears a resemblance to the distinction we made in the introduction between noematic, or objective, consciousness and noetic, or subjective, consciousness. The two approaches to understanding consciousness, according to the philosopher, are indeed introspection and scientific observation. While introspection gives "immediate" access to the content of consciousness, it operates within a field of intelligibility that is, so to speak, indifferent to the body. Indeed, it may be the case that no body exists, and introspection, in principle, could not less theoretically give us access to something. It could also make us perceive things that have no counterpart in what we commonly call reality, just as the dreamer does. On the other hand, scientific observation describes the functioning of bodies, at least as we represent them in our theories, but without providing the contents of consciousness that are mobilised. In other words, consciousness is not an observable element within the brain (assuming it is in the brain), as would be an object, a colour, a texture, etc.

9.4.3 Main Objections

Although McGinn's general idea follows the same line as the internalist and externalist chiasms described in the introduction, some criticisms could be raised. To begin with, it is naïve to think that introspection would provide direct access to consciousness. It would be more plausible to say that it certainly gives access to content, but that this access itself is conditioned by all sorts of factors external to introspection as such. I am not referring to the fact of having a brain, as this would commit a petition of principle. In the state of introspection, the thinker searching for a solution to the *Mind-Body problem* cannot affirm or deny the idea that, without their brain, introspection would be impossible; for the brain is a reality that exists beyond the senses. Nevertheless, introspection is carried out through words of language, conceptual and sensory categories, which do not seem to be the product of the individual themselves. Therefore, within introspection, there is reference to something other than our own consciousness. This fundamental otherness is what allows us to question methodological individualism: it could be that

[20]For a review of this argument, see for example Stanford Howdyshell, *On Realism and the Pessimistic Meta-Induction*, Kent State University, Critical Reflections, Essay on Significance and Critical Reflections.

introspection is not a faculty exclusively endowed to the human individual, but rather the result of a set of socially determined constructs. From this perspective, all introspection is mediated by language, concepts, and, more generally, by our culture.

Another critique concerns, this time, the idea that scientific observation would provide access to the body. Indeed, since this observation always occurs from within the human mind, it is not necessarily clear how it would not constitute, in turn, a sort of introspection, albeit mediated differently. A truly rigorous scientific stance would require stepping outside of our subjectivity and speaking, so to speak, from the point of view of God. In human practice, and this is precisely the starting point McGinn addresses, this is impossible. The distinction between introspection and observation seems, therefore, compromised. This is why we preferred to speak, in the introduction, of noème and noèse, of objectivity and subjectivity. Introspection is observation applied to states of thought. Observation is introspection that assumes the ideas to which this introspection refers exist in the world.

That being said, it is true that the brain can be conceived of as an object in the world. It is enough to abstract, when we represent the brain, from the act of perception itself, in order to retain only the idea of a brain existing independently of us. It is also remarkable that humans have the ability to conceive that things which exist in their mind also exist outside of them — and vice versa. This conceptual boundary between the object in itself and the object for us is probably one of the most essential properties of the mind. Without it, there is no doubt that most metaphysical problems would not exist! It is this boundary that makes it so that, when I consider the brain as an object in itself, I abstract from my perception. There is no longer, then, anything in this brain that belongs to the realm of perception. In this precise sense, one can say that scientific observation will never find, in the objects thus considered, anything related to consciousness.

We therefore think that the antinomy of the Mind-Body problem arises from the fact that there is a contradiction in the terms. Observation, by definition, sheds the object for itself of perception; and introspection, on its part, abstracts from the object in itself of observation. So that the real mystery, if there is a mystery, consists in knowing why we are capable of distinguishing the in itself from the for us. It goes without saying that certain ingredients can already be highlighted. First of all, it seems that this capacity to make the distinction is co-dependent on the ability to be conscious of oneself as a Cartesian "I think." For, indeed, if there were no in itself and for us, there would be no difference between what we think of the world, as consciousness, and what we think exists beyond our thought. It would therefore not be permissible, either, to think of oneself in the world or to think of the world within us. Philosophy is entrusted with the almost sacred task of undertaking the identification of the conditions that make this thought operation *possible*. Such an approach will be undertaken in the

final part of this thesis, which will seek to reconstruct the phenomenological concept of the "I". In the immediate term, I hope to have provided, through these few remarks, an interesting revision of McGinn's position.

9.4.4 Is Intuition a Mystery?

Before concluding, it is legitimate to ask how mysteriansim, which concerns human consciousness, offers an answer to the question of intuition. The question of solving the Body-Mind problem is, in essence, not quite equivalent to the question of drawing a line of demarcation between human thought and machine thought. This latter question, in any case, although it supports the metaphysical question, can be understood in a more functionalist sense: independently of any explanation of consciousness per se, what are the thought functions currently inaccessible to machines in existing paradigms? It happens that certain recent developments of mysticism offer an interesting insight.

In an article entitled *Level-headed mysterianism and artificial experience*, Prinz proposes to define the mystery of consciousness by means of an analogy: how many spoonfuls of cyanide does it take to become poisoned[21]? We know for a fact that one milligram is not enough. Certainly, that amount is necessary, but it will not suffice. It is thus a necessary condition, but not a sufficient one. Conversely, unless we are the devilish resurrection of Mithridates, we know that a ton is more than enough, but it is not necessary to ingest that much to pass *ad patrès*. This, then, is a sufficient condition, but not a necessary one.

From this example, it is easy to grasp Prinz's idea: isn't consciousness that unknown dosage? This revisited mysticism consists in saying that, regardless of the method used, it is impossible to determine all the necessary and sufficient conditions for the presence of consciousness in a given organism. This means that, whenever a property is considered, either it is a sufficient condition (it is sufficient to be a phenomenal *I* to have consciousness), or it is a necessary condition (it is necessary to be a living organism, to have memory, perceptions...), but never both at the same time.

Beyond the relief this strategy induces, like when a drowned person suddenly comes back to life, I propose the following idea: the concept of intuition contains at least one of the X elements of consciousness that renders it mysterious. In other words, this X is at the intersection of intuition and consciousness. The mysteriansim posture will be to say that it is impossible to give the necessary and sufficient conditions for having X, so that we cannot provide the conditions for having intuition or for having consciousness.

From this point on, and we arrive at the question we posed earlier, whenever intuition captures this X, it could not be explained through language.

[21] J. Prinz, *Level-headed mysterianism and artificial experience*, Journal of Consciousness Studies 10 (4-5):111-132, 2003

What I mean by this is that there are two possibilities. Either intuition is defined in such a way that it does not express the X property that is the subject of all the mysteries, or it expresses this X. In the first case, we can provide necessary and sufficient conditions. Provided these conditions are logical processes mimicked by a Turing machine, we will eventually construct a computer capable of intuition. The moment we wish to say something about X, it will not be possible to provide a necessary and sufficient definition of intuition.

Note that even if machines were capable of imitating all the functions usually attributed to human intelligence, they would not necessarily converge towards the mysterious X. We must certainly admit that all the functions of this machine are necessary to human intelligence, but this will not make these conditions sufficient. This scenario is analogous to the case in mathematics of convergent sequences that are not defined at their point of convergence.

Let us illustrate this geometrically by assuming that we ask a child to construct a circle by drawing segments of equal length on a sheet of paper. Furthermore, let us consider that the child, although they fully understand the definitions of the terms in the question, has no idea how to respond. This child is in a situation analogous to the five-year-old described by McGinn, who would attempt to conceptualise the theory of relativity.

The child can nonetheless realise that if they use three segments, they will obtain a triangle. If they use four, they will get a square; and the square is already closer to a circle. Continuing with this impossible quadrature, they will observe that a regular polygon with n sides is closer to the circle than a regular polygon with fewer sides. As n increases, and provided they are a bit clever, they might even prove that the sum of the lengths of the sides of the polygon, divided by the distance from the centre to any of the vertices, approaches π (but is always greater). Let us denote this ratio as P. It is thus necessary, for something to be a circle, that P be at least smaller than the one found in a regular polygon with n sides. The problem is that, no matter how large n is, P will never be equal to π. We have therefore generated an infinite number of necessary but insufficient conditions.

We conclude that, for a being who knows only a strictly finite number of segments, the concept of being a circle is unintelligible. This is not merely a mathematical puzzle, but it is also a transcendence of thought. This transcendence, however, is present in their mind. It exists, so to speak, tendentially — or asymptotically. The circle, far from being merely a mystery in the mind of our child, as it was before they explored the question, has become something properly mystical. For the Pythagoreans, the circle was, in fact, the perfect form. Thus, from this perspective, we could say that the mysterious X which forms the essence of consciousness and human intuition, this metaphysical *Omphalos*, this neuralgic centre to which all philosophical questions seem to converge, is something we can only grasp tangentially, but

will never be able to fully articulate.

9.5 Second solution : Intellectual Intuition

Up to this point, we have spoken of the dilemma of explicitation as though it presented only two opposing terms: either intuition can be made explicit, and is therefore reducible to a set of intelligible operations, or it cannot. In this section, I would like to suggest a third path: the expression of intuition within a theory that is not mathematical and thus, *de facto*, not reducible to automatic deduction — and yet which would remain fully explicit in language. This is a stance that can be traced back to Plato. It has been given the name — oxymoronic as it may seem — of *intellectual intuition*. I shall define it by turning to the original Greek text, after which I will examine the criticisms levelled against it by Kant. I will conclude by outlining a strategy for resolving the dilemma of explicitation.

9.5.1 The Contemplation of Ideas

Plato lays the foundations of intellectual intuition in a passage from the *Republic*, which I have already examined at the beginning of this thesis: the analogy of the Line. At the risk of repeating myself, I believe it is important to show how this text reveals the idea that, for the Ancients, there exists an order of truth distinct from simple deductive logic. This will also provide an opportunity to understand why Kant's critique, by reducing mathematical truths to the level of *a priori* synthetic judgements of the Understanding, made this path impracticable[22].

In the Analogy of the Line, we see Socrates conversing with Glaucon about the necessity for the city to be governed by philosophers and about the virtues of these philosophers. He compares the political organisation to a ship whose captain is indeed large and strong, but half-blind, half-deaf, and lethargic. The sailors fight for the helm, lull the vigilance of their leader with mandrake, and if one of them seizes the wheel, the others kill him and throw him overboard. In this context, the sage, the one who truly knows the art of navigation, is cast out from society. Yet he is the only one capable of "seeing," but his good nature, if uncorrupted, appears to others as the most eccentric.

The temperament of a sage, Socrates specifies, is like a plant growing in a foreign climate. It requires more care not to wither. And this plant, which is larger, can bear the finest fruits, just as it can produce the worst of them. In both good and evil, it will never suffer from mediocrity. That

[22]This repetition that I allow myself here, taking Plato and Kant as anchoring points, is also justified by the fact that I try to give as much self-sufficiency as possible to the different parts of my thesis. My ambition is indeed to have it published in the form of several essays, complementary but individual.

is why the virtue most necessary to him is the ability to conceive the idea of the good. This idea, however, does not belong to the sensible world. It is not the multitude of good things considered separately, just as the beautiful is not the multitude of beautiful paintings. It is much more like the sun; a source of light, nothing can be seen clearly without it. Remove it, and our philosopher is suddenly blind to the truth. But just like the sun, it cannot be stared at directly with wide-open eyes. The only way to ascend to this idea of the good is to start from hypotheses and, reasoning step by step, to rise to the idea beyond all hypotheses, which is called the anhypothetical.

In this respect, the philosopher's approach can be likened to that of the geometer. The disciples of Euclid certainly speak of triangles, squares, and diagonals, but what matters to them are not the sensible forms drawn in the sand to help them reason: it is the mathematical ideas themselves. However, there is an essential difference between the geometers and the highest intelligence, both in Plato's thought and, as was the case with Pascal. The geometers, in fact, only see the principles and rely on images to reason about them. They proceed through analysis and deduction. The sage, however, when contemplating the idea of the good, has no sensible copies at his disposal. He cannot draw the good in the sand and deduce something from the essential idea. He must formulate hypotheses beyond the visible representation. Thus, the contemplation of the idea of the good is not the result of deduction or analysis, but of contemplation, synthesis, or even intellectual intuition.

Truth, according to Socrates, can therefore be represented along a line divided into four equal segments. The first two segments represent the two levels of reality in the sensible world. The last two segments, for their part, represent the two levels of reality in the intelligible world: that of logical-deductive reasoning (*Diánoia*) and that of the Forms (*noèton*). I would like to recall below how Socrates contrasts these logical-deductive reasonings, which fall under simple *Diánoia* calculation, with the faculty of intuitively contemplating higher truths (*Noèsis*). Indeed, at the turn of these pages, there is an inaugural assertion: the highest faculty of man would not consist of calculating like today's machines, but of contemplating knowledge, feeling it, immersing oneself in it with ecstasy and fascination, in the way Galileo sought to gaze at the Sun:

Noèton *Epistémé* (science)		**Horatón** *Doxa* (opinion)	
Noèsis (intuitive)	*Diánoia* (discursive)	*Pistis* (belief)	*Eikasia* (imagination)
↓	↓	↓	↓
concerns the anhypothetical	concerns the hypothetical	concerns the sensible objects	concerns their images

This notion of contemplation in Plato, though essential to his thought, is undoubtedly an inheritance of Pythagoreanism. Far from being confined to a purely rational or symbolic theoretical teaching, the Pythagoreans combined the teaching of mathematics, music, and the philosophy of nature with an esoteric teaching. They encouraged a communal life, reserved for initiates, but one with a political project for the whole of society. The one who wished to attain the truth from Master Pythagoras was not merely faced with a theoretical discourse. He was not just to strive to adopt the right method and free himself from prejudices. His learning also required him to sign the death of the profane self he once was, to surrender to an inner transformation, to change his outlook on the world. One could say with Foucault that truth, as thus conceived, "is only given to the subject at a price that engages the very being of the subject. For as he is, he is incapable of truth[23]. "

The transformation of oneself through the knowledge of first principles is, of course, not foreign to Plato. It is plausible to think that it is part of the definition of intellectual intuition that he gives in the *Republic*. What is at stake, indeed, is the existence of a faculty that allows one to contemplate the truth beyond science — science being but the stepping stone to a higher reality. This idea is beautifully explained in this passage from Socrates:

> *What is, then, of all the gods of the sky, the one [...] whose light allows our eyes to see in the best possible way? [...] It is the sun, of course, that you ask me to name. [...] Just as one is right to think that light and sight are similar to the sun, but wrong to believe that they are the sun, so, in the intelligible world, it is right to think that science and truth are each similar to the good, but wrong to believe that either one is the good*[1].

In Plato's thought, mathematical ideas, although abstract, do not hold the same status as the ideas of the Good, the Beautiful, or the True. The former indeed belong to the realm of what is deducible, they are principles or hypotheses. The latter belong to contemplation, they are anhypothetical or unconditional truths. In other words, contemplative intuition is held to be superior to logic.

The recent history of philosophy has, as it were, inverted this state of affairs, making mathematics, if not the foundation of all possible knowledge, at least the model of scientificity. This attitude is aptly captured in the expression "The Queen of Sciences," which is attributed to mathematics, leaving the philosopher-kings to debate their eviction. However, it is interesting to ask how we came to abandon this idea of intellectual intuition.

[23] Foucault [?], *L'Herméneutique du Sujet*, p. 9.

9.5.2 Kantian Critique

Among the authors who have dealt a fatal blow to intellectual intuition, Kant's role must not be overlooked. In the *Critique of Pure Reason*, the typology of human faculties—ranging from sensibility to reason, via the Understanding—follows, in a sense, the analogy of Plato's divided line, but does so while dispensing with the fourth term. It is not that *dianoia* has strictly disappeared with the advent of the Kantian transcendental subject, but it is as if the anthropological depth of the Platonic being had been flattened out. The concept of Reason, in Kant's view, is the faculty through which the mind seeks the unconditional principles upon which the categories of the Understanding rest. Reason is thus analogous to Plato's *dianoia*, with the important difference that it operates not through any kind of contemplation of truth, but through the mediation of pure reasoning. Yet this pure reasoning, Kant tells us, because it cannot rely on any criterion of truth external to itself, leads only to antinomies.

Let us clarify this idea. In Plato, the Ideas are the objects of thought par excellence. These Ideas exist independently of the human mind. They must therefore already be given in order to be contemplated. Under these conditions, it becomes difficult to defend the possibility of strictly *a priori* knowledge, such as that of mathematics. This is why, according to Kant, the true method in speculative philosophy is to assume that it is the objects that conform to human knowledge, and not the reverse. This conceptual reversal is a genuine Copernican revolution: "If intuition had to conform to the nature of the objects, I do not see how we could know anything about them *a priori*; but if, on the contrary, the object (as an object of the senses) conforms to the nature of our intuitive faculty, then I can very well understand how such knowledge is possible[24]." Everything that is the object of *a priori* knowledge is always already within us, prior to all sensible or intellectual experience. With Kant, intellectual (noetic) intuition is replaced by the *a priori* synthetic judgement.

A major consequence of the Kantian turn is that it is no longer possible to know things except as phenomena. Indeed, insofar as our concepts conform to the objects, and not the reverse, we never know objects except in the modality of our perception. As a result, it becomes impossible to extend knowledge to anything other than objects of a possible experience. Yet within the human being there lies a natural disposition, in the ongoing development of knowledge, to ask questions that transcend all intuitive representation. These are the questions concerning God, the world, and the soul. Is the world finite or infinite? Did it have a beginning in time? Was it created by a Supreme Being? Is the soul a simple and incorruptible substance? These are the questions that have nourished all systems of belief, whether metaphysical or cosmogonic. Since such ideas refer neither to

[24] KANT, *Critique of Pure Reason*, trans. BARNI, Preface to the Second Edition, p. 24.

pure intuition nor to any sensible experience, they always lead to paradoxical reasoning and, for this reason, must be relinquished. The purpose of the *transcendental dialectic* is to examine with great care each of the logical conclusions to which such considerations lead, in order to demonstrate their inextricably aporetic nature.

Upon close inspection, the first result of this critique undertaken by the philosopher of Königsberg is, all things considered, a *negative* one: it seeks to demonstrate that the use of pure reason cannot lead to properly scientific knowledge. Kant thus proscribes Plato's *noēsis*, conceived as the search for the anhypothetical. Yet this restriction, far from dethroning dialectic from the pedestal upon which Plato's philosophy had placed it, contributes on the contrary to confirming its position. To be sure, it must now descend a few steps, and abandon its claim to encompass the totality of knowledge at a glance; but the exercise of dialectic as such is preserved: "There has always been, and will always be, a metaphysics in the world; but there will also always arise alongside it a dialectic of pure reason, for this dialectic is natural to it. The first and most important task of philosophy is therefore to deprive this dialectic, once and for all, of all pernicious influence by destroying the source of its errors[25]." It therefore seems more judicious, rather than going against Kantian critique, to ask whether, in our scientifically inscribed world, it is still possible to justify the existence of a faculty comparable to intellectual intuition. Such intuition would require a rigorous mathematical definition and be grounded entirely in deductive judgement, while nonetheless being wholly unreproducible by a machine. Though such an idea may appear audacious, it contains, upon serious consideration, nothing inherently contradictory. In our daily lives, we handle objects whose very nature escapes the paradigmatic workings of computing. As an example, one might think of analogue music—or, more generally, physical phenomena belonging to the order of the continuum—which must be discretised (or digitised) in order to be recorded and modelled by a computer. Could it not be that the human mind, by virtue of its intuitive dimension, is capable of grasping the continuum?

It is this idea, presented here in embryonic form, that we shall develop in the chapters to come. It appears, in the face of the dilemma of intuition, as a third path. Rather than maintaining that intuition is undefinable in language, or that it is only definable through a theory that lies either beneath or beyond mathematics, we shall advance the following claim: *the concept of intuition can be explicated through a language which, though mathematical, is not implementable by a machine.*

This position will lead us first to define how a computer functions, by setting forth the principles of the Turing machine and the Church–Turing thesis, as well as the principal limitations to the computational power of

[25] KANT, *op. cit.*, Preface to the Second Edition, p. 36.

the computer. Once the mathematical definitions of thought automation are laid out, it will be necessary to ask what shortcomings these procedures reveal—or, in other words, to determine which kinds of operations, though still mathematical, these machines are unable to carry out. Is it possible that such operations are precisely what defines human intelligence? Failing that, can we assert with any certainty that intuition is reproducible by a machine?

What to take away from this chapter

Attempting to define why the machine lacks intuition gives birth to a dilemma historically identified by Alan Turing:

- Either one defines intuition explicitly by means of operational rules, but in this case, it is unclear why a machine could not learn those rules.
- Or one refrains from defining intuition in explicitly operational terms, but then one runs the risk of no longer knowing what intuition truly is.

In this chapter, I discussed two approaches aimed at circumventing this dilemma:

1. Within *mysterianism*, one assumes that an operational definition exists, but that it is cognitively closed: it is therefore impossible for us, as humans, to access it.
2. Within *mysticism*, an explicit definition exists, but it is not expressed in the language of reason. At most, it is a symbolic definition.

The purpose of the next chapter will be to explore a third path: to provide intuition with an explicitly operational definition, such that it cannot, however, be implemented by machines. In technical terms, we shall say that the function must be mathematically *well-founded*, yet *non-computable*.

Chapter 10

Chuck Norris's Machine

> Chuck Norris already computed at infinitum. Twice.
>
> Anonymous

We have seen that every attempt to explain what intuition is runs up against a major dilemma: either the explanation is sufficiently clear and operational to be described by a mathematical function, in which case there seems to be no principled objection to a Turing machine being able to implement it; or, on the contrary, the explanation suffers from ambiguity, such that one cannot be sure what exactly is being described. I have shown that two of the strategies traditionally considered by philosophy to circumvent this problem encounter significant objections. Neither mysteriansim nor the hypothesis of intellectual intuition appears to offer a satisfying answer.

In this chapter, I would like to explore a third path: to provide a definition of human intuition which, although mathematically explicit, cannot be implemented by a computer. This answer thus amounts to the claim that the human mind is capable of performing mathematical operations that nevertheless elude the computational power of the Turing machines we have examined so far (see 5.3.1).

As we have said, it is crucial that my definition of intuition be expressible within a mathematical formalism. Yet it seems difficult to mathematise a notion as vague as intuition without making scientific assumptions about how the mind works. If I were to say, for instance, that intuition consists of a set of information processed by the brain without being consciously accessed, this would presuppose, on the one hand, that the human mind can be represented as a machine capable of processing information, and on the other hand, that the mind operates through the brain. Such a vision is easily compatible with the computationalist paradigm, but it could be challenged by proponents of enactivism or the extended mind thesis, who seek to explain cognition not only through the brain but also through the

body and its relationship with the environment.

To circumvent this difficulty, I will not attempt to build a mathematical model of human intuition, but rather a model of the mathematical functions of human intuition. What distinction exactly do I have in mind when I speak of models and functions? I believe everyone agrees with the idea that human beings can carry out calculations on numbers. Addition, subtraction, multiplication, and division are mathematically well-defined functions which can be performed by a human mind, an abacus, or a calculator. This does not, however, mean that the human mind operates in the same way as an abacus or a calculating machine. Terminologically, we would say that addition is a function, while the human mind, the abacus, and the computer are models capable of executing this function.

When it is said that the human mind behaves like a computer, it is often meant to express nothing more than this: the functions the human mind can execute are also, to a large extent, satisfied by our computers. Nonetheless, notable operational differences remain: whereas the human mind is composed of glial cells, myelin, and nodes of Ranvier, the computer is woven from wires and epoxy boards, transistors, capacitors, and microprocessors. Where the computer executes sequential, binary operations governed by an internal clock, the brain activates, in parallel and by means of continuous signals, a vast number of interconnected cerebral regions.

The totality of functions a computer can perform is precisely defined by the formalism of the Turing machine. A computer is, in essence, a logical system capable of performing calculations. The great idea of Alan Turing and Alonzo Church, known as the Church-Turing thesis, states that every computable function can be executed by a Turing machine (see Appendix 3). In other words, all that a Turing machine can do is compute. The question that arises is whether the human mind is capable of more than computation. Or, to put it another way, whether there exist mathematical functions that appear to elude the formalism of the Turing machine while nevertheless remaining within reach of human intelligence?

In this chapter, I will begin by drawing a sharper distinction between a model of the human mind and a function of the human mind. I will also anticipate certain objections that will be addressed in greater depth in the following part of the thesis. Finally, I will suggest three strategic approaches that may allow us to construct a function that exceeds the computational capacities of the Turing machine.

10.1 Functions and Models

To begin with, it seems crucial to clearly distinguish what is meant by *function* and by *model*. A *function* is a relation that associates, to any given value, a corresponding output. It resembles the Sibyl of ancient Rome. You

ask her a question; she provides the answer. This is, in fact, the very meaning of the mathematical definition. Students learn as early as secondary school that a function is a transformation which takes a given object from the domain of definition (for instance, a natural number) and returns, for each such object, another object from the range (for example, another natural number). Nothing, however, forbids this function from operating not with numbers, but with fruits and vegetables. It may be a transformation that tells you which vegetable to buy according to the season, for example:

$$\begin{array}{rcc} spring & \rightarrow & avocato \\ sumer & \rightarrow & strawberry \\ fall & \rightarrow & peer \\ winter & \rightarrow & lemon \end{array}$$

The *model*, if I may return to the mythological metaphor, is a description that seeks to explain what happens inside the Sibyl's mind when she gives an answer to a certain question. The model attempts to understand where this intuition, this visionary art, comes from, allowing her to foresee the future. It is, in other words, an explanatory theory that rationally elucidates the link between the inputs and outputs of a given function.

To summarise what has just been said: the function can be described as a black box linking an input to an output, whereas the model is a mechanism, equipped with operational rules, capable of executing one or more functions. This distinction allows us to understand that the computer and the human mind may indeed share a common function, namely that of computing, without this implying that the computer constitutes a valid model of the human mind. What I mean by this is that if you give an arithmetic operation to someone skilled in mental calculation, they will — or so we hope — produce the same answer as the machine. The same equation always leads to the same result. It is, however, doubtful to suppose that the human mind and the computer exhibit identical internal operating mechanisms. In other words, the way in which the answer is produced does not necessarily follow the same rules.

What are the exact differences in operation between the computer model, which is none other than the Turing machine (see section 5.3.1), and the model of the human mind? Let us observe that the human being, unlike the computer, is liable to make frequent reasoning errors — particularly when the calculation exceeds their working memory. Furthermore, it is not certain that "computing" is the most essential function of the human mind, whereas it is unquestionably the primary function of the computer. There are, in fact, cultures in which the very concept of number is absent. There are also individuals who suffer from dyscalculia. Whether you are dyscalculic or come from a pre-literate tribe, that does not prevent you from possessing a human mind capable of reasoning, of social cognition, of adaptation, and

so on. Ultimately, it may be that computation, understood as the capacity to carry out strictly logical-deductive reasoning, represents only one facet — and a rather secondary one at that — of human aptitudes.

In what follows, I would like to ask in what precise ways the human mind and the machine, considered as models of computation, may differ.

10.1.1 SNARC Effect: The Intuitive Structure of Arithmetic

How does human calculation truly work? What elements does it rely on that set it apart from the machine? If you ask those around you, you may find that some people perform mental arithmetic without being able to say where the answer comes from. For them, the result emerges as though by unconscious necessity. Others, by contrast, visualise digits in their mind's eye and manipulate them mentally. Still others possess the curious ability to represent numbers within visual or peripersonal space. For them, numbers occupy specific positions in space; they float in the air and can be pointed at with a finger. Finally, there are those who see colours. They associate each number with a particular hue, or even with a sound, or a character. Arithmetic operations, for them, are colourful frescoes, musical phrases, sensory experiences.

Despite the diversity of these experiences, can we identify a common model to explain how the human mind carries out arithmetic operations? Recent studies suggest the existence of an implicit visuo-spatial structure, wherein numbers are arranged along a line stretching from left to right (at least in Western cultures[1]). To test this hypothesis, an experimental set-up was devised. Within it, each participant is presented with two numbers between 0 and 9; one appears on the left side of the screen, the other on the right. The task is to determine which of the two numbers is greater. Interestingly, it was observed that response times were shorter when the larger number appeared on the right. This bias is what is known as the SNARC effect.

The SNARC effect suggests that there is an automatic — one might even say natural — tendency to associate numbers with positions in space. This spatial visualisation is conscious for some individuals; for others, it remains unconscious. Synaesthetes, for example, are capable of drawing the position of numbers in space. It is even impossible for them to think of numbers without this spatial scaffolding. Figure 10.1 offers an illustration of how a synaesthete might represent the scale of natural numbers. One notices that the numbers are oriented from left to right, consistent with the results supported by SNARC effect studies. Yet this orientation is far from linear: one observes convolutions that seem to defy the imagination[2].

[1] Dehaene et al., *The mental representation of parity and number magnitude*, J. Exp. Psychol., 122, 371–396, 1993

[2] In this example, each number, in addition to occupying a spatial position, is also

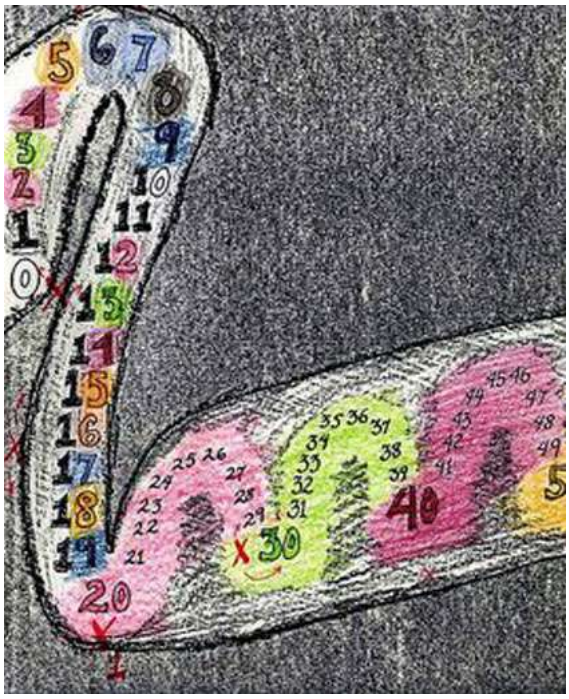


Figure 10.1: Representation given by a child exhibiting spatial–numerical synaesthesia.

It is reasonable to believe, as a hypothesis, that this spatial visualisation of numbers enables human beings to reason about infinite magnitudes. When one speaks of the sequence of powers of two, for example, one envisions a spatial magnitude that is always twice as large as the previous, ultimately tending asymptotically towards the notion of an infinite line. The possibility of such a line is visually clear, even though it cannot be grasped in its full extent. One might say that the spatial representation of numbers constitutes an *a priori* perceptual structure, within which our intuitive understanding of arithmetic operations takes place. This structure exhibits a well-founded order. There are numbers that come "before", and others that come "after". There is also an initial point (zero) — but no limit to the process, which stretches endlessly toward horizon[3].

A computer does not appear to possess such an intuitive structure. It

associated with a colour. The association between a number and a colour is called grapheme–colour synaesthesia, while the spatial positioning of numbers is known as spatial–numerical synaesthesia. The two do not always coincide.

[3]This notion that there exists an intuitive structure of perception was already discussed in Chapter 9. There, I defended the idea that, unlike a computer program, the human being is subject to optical illusions. Intuition can be conceived as automatic associations influencing judgement — structures that link concepts to one another independently of our will.

does not automatically associate numbers with a spatial magnitude. The proof lies in the fact that a computer is not susceptible to the SNARC effect. There is no reason to think that, if one presents a computer with numbers on a sheet of paper, the position of those numbers relative to one another would influence its speed of execution in the same way it does for humans. What I mean is that one can write a computer program that bypasses the SNARC effect, whereas it is impossible for a human being not to be subject to it.

The absence of an intrinsic spatial visualisation of numbers might explain why the machine is incapable of assigning meaning to infinity. A computer can indeed manipulate numbers that tend asymptotically toward infinity, but can it represent the infinite line towards which this sequence of numbers tends? What is specific to the human visualisation of a line is that the notions of actual infinity and potential infinity meet, as it were, automatically. That is to say, the human mind becomes capable of conceiving that a sequence of increasing numbers converges towards something it names infinity. It imagines a concept — that of an infinite line — which lies beyond all possible representation. It gives visual form to a concept that remains, in itself, abstract.

From these few reflections, two conclusions may be drawn. First: the calculator is not necessarily a model of the human mind, any more than the Turing machine is a model of the brain. What human beings and machines have in common is not their mode of processing information, but rather their ability to perform a certain number of functions — calculation among them.

Secondly — and this is the more important point — it may be that the workings of the human mind allow it to implement functions that exist neither in a calculator nor in a Turing machine. We have pointed out, notably, the ability to visualise infinity, or to understand certain properties of sequences tending towards infinity. It may therefore be that, by exploring this aspect further, we shall discover that there exists a function within the human mind that requires a model other than the classical Turing machine. This is why, in the following section, I shall turn to the question of what a Turing machine might look like if it were endowed with the capacity to represent infinity.

10.1.2 Towards a Model of Transfinite Computation

The history of the philosophy of artificial intelligence is strewn with declarations—concerning intuition—that later proved inaccurate. Hubert Dreyfus once claimed that no machine could defeat a human at chess, on the grounds that it lacked the intuitive learning capacity necessary to instantly recog-

nise patterns[4]. The thinkers of the Palo Alto school, for their part, asserted that a machine could never navigate a city, as such a task demanded a level of contextual understanding far too subtle. These examples, now obsolete, demonstrate that computational methods have time and again refuted objections grounded in intuition.

Yet is it still possible to raise a principled objection? Might one construct an argument that stands independently of the contingent limitations of contemporary computing? This would entail demonstrating, with rigorous clarity, that the kind of intuition under discussion (learning, contextual sensitivity, etc.) is inimitable by a Turing machine. Several approaches to this task present themselves:

- **Strategy based on undecidability**: A first approach involves defining intuition in such a way that the function it performs is mathematically reducible to an undecidable problem. Among such problems lies the question of whether a given Diophantine equation has a solution. A Diophantine equation is a polynomial equation for which only whole-number solutions (integers) are allowed. Yuri Matiyasevich's theorem has indeed shown that it is impossible to construct an algorithm, executable on a Turing machine, that could answer this question for any Diophantine equation whatsoever. Algorithms can be devised for particular cases only. If, therefore, we define intuition as "the capacity to determine, in the most general case, whether a Diophantine equation admits a solution", we would then hold something that no machine could do. All that would remain would be to argue that the human being, by contrast, can. This type of strategy has been popularised by the Lucas–Penrose argument, which hinges on a similar trick—albeit one that relies on Gödel's theorems rather than on Turing machines. One of the principal objections such strategies must contend with is the fallibility of intuition. Intuition is rarely depicted as an ability suited to every scenario. It is prone to error, but also to silence. It thus corresponds only loosely to "the capacity to determine *with certainty*, *in the most general case*, whether a Diophantine equation admits a solution."

- **Strategy based on randomness**: If one takes seriously the random, dazzling, and capricious nature of intuition, one might be tempted to model it using chance itself. Randomised algorithms are not part of the toolkit of a classical Turing machine. One must instead appeal either to non-deterministic Turing machines or to quantum computers. Certain quantum algorithms do in fact show that randomness can enhance the resolution of problems belonging to the class of NP-complete

[4] Dreyfus, *What computer still can't do ?*

problems, widely considered among the most difficult in mathematics (see section 6.6). Solving such problems by brute force—verifying each possible solution one by one—requires exponential time in the worst case. A quantum algorithm reduces this time by a square root. This is still exponential, but with a meaningful speedup. Might human intuition operate differently? The central issue with this line of reasoning is that one might respond by saying that intuition is merely a matter of computational paradigm. Once quantum computers are in our hands, there will no longer be any reason to think of intuition as a uniquely human faculty.

- **Strategy based on infinity**: This is why the most general argumentative line we may envisage consists in claiming that intuition—and particularly the intuition of the infinite, discussed earlier—requires the ability to compute the infinite. What do I mean by this? That it be possible to reach the limit of an infinite process, such as the enumeration $1, 2, 3, \ldots$ This capacity lies beyond the power of any Turing machine, and arguably, beyond any machine that might be physically constructed. Yet it is possible to build an extension of the Turing machine: the transfinite Turing machine. This is the strategy I shall pursue in this thesis.

The task before me is twofold. On the one hand, (1) I must show that the three functions just defined are precisely those enabled by the intuition of the Infinite and of the Whole, which I have earlier identified as the fundamental ingredients of human perception. On the other hand, (2) I must also construct a model of these functions. This model will take the form of a transfinite extension of the classical Turing machine—that is, a machine capable, like Chuck Norris, of counting to infinity... several times in a row.

This programme, whose nature I have just unveiled, is more ambitious than it may initially appear. To construct a model that unites both types of intuition (that of the Whole and that of the Infinite) and all three functions (undecidability, randomness, and infinity) amounts to demonstrating the existence of a conceptual equivalence between these disparate notions. More precisely, it supposes the existence of a mathematical unity X that links undecidable problems, NP-complete problems, and infinite series, and then constructing explicitly a model that realises X. Far from belonging solely to the realm of philosophy, this project might also yield insights of interest to mathematics. I do not claim to prove new results, but to offer, on well-known problems, a different perspective. This enigmatic X is precisely the transfinite Turing machine.

Before launching into a formal definition of the transfinite Turing machine and examining its chief properties, I would like to begin by justifying

the claim that the three functions I have just enumerated do indeed capture the intuition of the Infinite and the intuition of the Whole. More specifically, I aim to show that the intuition of the Infinite is linked to two functions: the completion of infinite series and the resolution of undecidable problems; while the intuition of the Whole is linked to the resolution of NP-complete problems (see table below).

Definitions	Functions
Transcendent intuition	Completing infinite series Solving undecidable problems
Holistic intuition	Solving NP-complete problems

10.1.3 The Infinite and Undecidable Problems

It is easy to conceive that the intuition of the infinite enables one to grasp the convergence of infinite series. This, at any rate, was the case I sought to make when discussing synaesthesia and the SNARC effect. It is not, however, particularly evident why the resolution of undecidable problems or NP-complete problems would be related, in any way whatsoever, to this intuition of the infinite.

Here, I would like to offer a clarification of this point—one that shall later receive more ample justification. The claim will be as follows: that the resolution of undecidable problems is linked to the intuition of the Infinite, whereas the resolution of NP-complete problems is tied to the intuition of the Whole.

Let me begin by suggesting that a machine capable of counting the infinite would no doubt also be able to solve undecidable problems. In order to grasp the link between infinity and undecidability, it may be wise to recall a few mathematical results. When the first undecidable problems were discovered by Gödel and Turing, logicians turned their attention to the question of how such problems might be circumvented. This led to a rigorous clarification of the assumptions underlying Gödel's theorems—assumptions that, in popular accounts of logic, are far too often overlooked. This is why I took care to recall them in Appendix 2.

One of these assumptions concerns us most intimately. A mathematician by the name of Gentzen reached the conclusion that, were one to possess a transfinite proof system, undecidabilities would be resolved. Put otherwise, undecidability arises from the fact that a mathematical proof must be of finite length. One begins with a set of hypotheses at the start of the reasoning, proceeds through a sequence of steps, and finally arrives at a conclusion. Now, it turns out that if one were allowed to carry a proof on into the infinite, it would become possible to give answers to undecidable problems. It would be possible, in particular, to prove the consistency of

arithmetic.

One need not be an expert in mathematical logic to understand that a Turing machine capable of counting the infinite would in all likelihood be able to solve undecidable problems. I shall prove this result in due course. For the time being, I would like to say just a few words more on the relationship between the intuition of the Whole and NP-complete problems.

10.1.4 The Whole and NP-Complete Problems

Unlike undecidable problems, the resolution of NP-complete problems does not appear to require the infinite. A well-known example of an NP-complete problem is the resolution of the computer game Minesweeper, which consists in clicking on squares while avoiding hidden bombs. One easily understands that this game presents a certain complexity, but it is difficult to see how a machine capable of counting to infinity would be more effective at playing it. The goal, after all, is to solve the game in finite time, so our ability to count the infinite would hardly advance us towards that end.

This is why I will argue that NP-complete problems are linked not to the intuition of the Infinite, but to the intuition of the Whole. There is indeed something in problems of the Minesweeper kind that demands the construction of an overarching view. The computational complexity arises from the fact that the information concerning a specific square on the game board exerts an influence on squares that may be far removed from it. Everything happens as if one had to keep in mind all the possible configurations of the game in order to produce an optimal response. Yet this overview quickly exceeds the computational abilities of a human being. The human player will therefore rely on experience. They will attempt to narrow the problem by focusing on a small number of squares, but whenever they lack the necessary information, they will turn to pattern recognition. A high-level Minesweeper player relies on the shape of patterns to guide their choices.

To properly understand the general structure of an NP-complete problem—and why it is tied to the intuition of the Whole—I will demonstrate that an NP-complete problem can be given a visual geometric illustration. More precisely, I will claim that an NP-complete problem consists in identifying, from a given graph, the points at which a function vanishes (see illustration 10.2). For human beings, there exists an effect of salience whereby our gaze is spontaneously drawn to these points. We are thus able to locate them without having to scan each and every x-coordinate. Now, unless proven otherwise, a Turing machine does not seem capable of determining quickly the points at which the function vanishes: the speed of the algorithms, in the most general case, is not significantly faster than if they were scanning the x-axis one coordinate at a time.

Returning to the Minesweeper example, even our best-known algorithms risk enumerating all possibilities one by one: they will consider the case

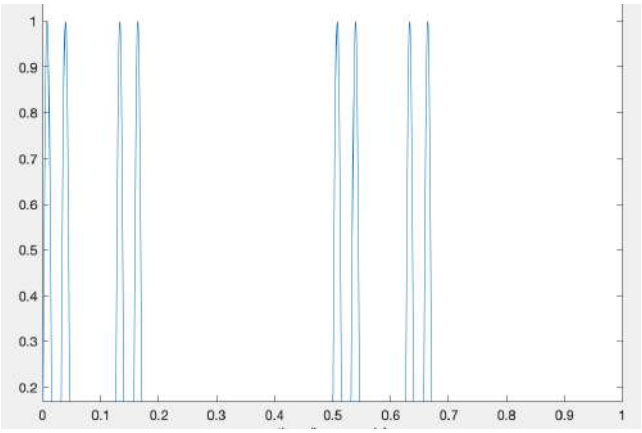


Figure 10.2: When given a graphical representation, an NP-complete problem consists in identifying the points at which a function vanishes. In the example above, there exist many solutions, but imagine a more complex diagram admitting only a single one. Provided that the graph has a resolution high enough for the human eye, I postulate that the eye is spontaneously drawn towards the "dips". It can therefore identify the presence of a solution in "fast" time, without needing to methodically scan the x-axis one coordinate at a time.

where a bomb lies on square A and square B, the case where a bomb is on square A but not on square B, the case where it is on square B but not on square A, and so on. The high-level human player cannot afford to examine each case individually. They will refer to their experience and instinct to immediately direct their attention towards the most promising scenarios—that is, not only the most probable ones, but also those that will allow them to solve the problem more swiftly.

This capacity for salience I speak of is not without resemblance to the workings of human perception discussed in the analysis of temporal illusions (chapter 3). Consciousness is always consciousness of something that changes, and for this reason, it never considers the points of a graph in isolation, in purely sequential fashion; rather, it sees them as sets that are partially distinct, partially comparable to one another. One could say that, under its gaze, the points of space are *entangled*. It does not perceive each scenario in isolation, but automatically associates one scenario with another. The information presents itself within a certain pre-structured form, which experience allows it to refine over time.

10.1.5 Responses to a Preliminary Objection

Before concluding this section, I would like to take a moment to anticipate a crucial objection. I have claimed that only a transfinite machine could

implement the functions of intuition. Yet it may seem problematic to assert that intuition is a faculty capable of grasping the infinite, when the human brain itself is firmly enclosed within the skull. How can an organ made up of a finite number of neurons, with a bounded volume and a limited processing speed, possibly apprehend the infinite?

This methodological objection admits of several responses. First of all, it is important to note that although the brain is finite, this is not necessarily the case for the mind itself; reducing the human mind to the computational capacities of neurons constitutes a petitio principii, since it presupposes that the mind is something computable. But this is precisely the question at stake. I will show, in the final part of this thesis, that the idea according to which the mind is located in the brain—however convincing it may appear—is in fact thoroughly naïve. My critique will rely on a philosophical reading of neuroscience.

That said, I am quite willing to accept the idea that the human mind is just as finite as the brain. It cannot, indeed, represent the infinite in *actu*—that is, the infinite in its totality. At most, it can form an asymptotic idea of it, imagining spatial expanses, numerical scales, or temporal magnitudes that grow without bound. It is thus capable of representing the infinite in *potentia*. When confronted with number sequences such as $1, 1, 2, 3, 5, 8, 13...$, the mind can (1) attempt to predict their regularities (could it be the Fibonacci sequence?), visualise that the series diverges toward infinity, and represent the end of the enumeration by means of the infinity symbol; it can even rank the different kinds of infinities, as in Cantor's theory, and imagine infinities of ever-greater magnitudes.

To conclude my case in favour of transfinite Turing machines, let us observe that many phenomena, bounded in time and space, are nevertheless modelled using infinite notions. Euclidean geometry, for instance, assumes that space is infinite—even though our universe has a clearly delimited size. The abstractions of infinite straight lines (infinitely large) and dimensionless points (infinitely small) are undoubtedly beyond the reach of our everyday experience of space, and yet they remain relevant for describing that very experience[5].

Another example is that of thermodynamics, which presupposes the existence of an open space. Such a space is theoretically possible only if the universe were infinite. Infinite spaces also appear in the formalism of Hilbert

[5] One might object that Euclidean geometry does not aim to model the Universe, and that it has largely been supplanted in contemporary physics by non-Euclidean geometries. According to the Einstein–Lemaître model, for instance, the Universe possesses a variable radius of curvature, oscillating between an elliptic and a hyperbolic space, only taking the form of a Euclidean space at a particular, entirely ephemeral moment t, when the curvature radius becomes null. This objection does not threaten the point being defended here, since hyperbolic space itself presupposes the infinite. Moreover, our experience of space on Earth is adequately modelled by Euclidean geometry (at least over short distances), even though the entire planet is more accurately represented by spherical geometry (which is finite).

space, used to describe the equations of quantum physics. Finally, let us not forget that the classical Turing machine is already an idealisation. Its memory and processing capacity are assumed to be infinitely extensible[6]. This does not necessarily correspond to the experience of the human mind, nor even to the actual computers we manipulate, which possess finite resources.

10.2 The Transfinite Turing Machine

In the preceding section, I limited myself to setting out a specification sheet. I outlined, point by point, certain properties that a machine ought to possess in order to fulfil the functions of intuition. The first criterion—one that immediately drew my attention—consists in building a machine capable of determining the convergence of an infinite series, even in the absence of known laws. Such a machine ought to be able to affirm that the sequence $1 + \frac{1}{2} + \frac{1}{4} + \ldots$ converges towards 2. It should also be capable of enumerating a sequence of alphabetical symbols of the form *abcaaabcaa* in order to determine whether this sequence "ends" in an infinite succession of *a*'s, or whether it loops endlessly on the pattern *abcaa*.

The second strategy, for its part, follows in the footsteps of the Lucas-Penrose argument. It involves constructing a machine capable of surpassing certain results of undecidability. These might include Gödel's incompleteness theorems, but also the halting problem, Rice's theorem, and so forth.

The third strategy, finally, consists in endowing the machine with a certain degree of randomness and hoping that this might account for differences in function concerning, for instance, the resolution of complex problems such as those falling under the P-NP classification. It is indeed often acknowledged that the ability to find, by "chance", a good candidate solution to this sort of problem typically stems from a form of intuition—whereas the machine proceeds only through a sequence of deterministic calculations.

In the pages that follow, I would like to propose a revision of the Turing machine capable of satisfying, at least partially, each of these three criteria. Although the model I intend to construct is primarily designed to "compute the infinite" (the first criterion), I shall show that it possesses very useful properties with respect to undecidability theorems (the second criterion) and the resolution of NP-complete problems (the third criterion). Put differently, I would like to build a formalism capable of capturing the notion of intuition in each of its three functional meanings.

My revision will consist in extending the classical definition of the Turing machine, both in its spatial and temporal dimensions. Regarding the spatial

[6] The Turing machine is equipped with a tape that can be extended as often as necessary. This tape is used to record symbols. It thus functions as both storage space and workspace. To express it through a metaphor, it is the sheet of paper upon which its calculations are inscribed.

dimension first, I shall extend the read-write tape so that it may contain an actually infinite number of symbols. As for the temporal dimension, I would like the machine to be able to perform an actually infinite number of operations. By "actually infinite", I mean that the machine can count the infinite and write $1, 2, 3, \ldots$ until it reaches the symbol ∞. It can then resume its reading to write a new sequence of numbers $1, 2, 3, \ldots$, and so on, recursively.

This approach I propose to undertake raises a number of difficulties. One must first ask what it means to count to infinity. It is indeed easy to observe that there are infinities greater than others. Enumerating the sequence of natural numbers $1, 2, 3, \ldots$ seems easier than counting the set of real numbers. There are indeed infinitely many real numbers between 1 and 2, but also between 1.5 and 2, or between 1.9 and 2. In fact, it is well established that, if you take any two real numbers, there will always be an infinity of real numbers between them.

This idea that the set of real numbers $\mathbb{R}$ forms a larger set than the set of natural numbers $\mathbb{N}$ was formalised by Georg Cantor at the end of the 19^{th} century (see Appendix 4). His great insight may be summarised in a thought experiment. Suppose you have two large sacks, each containing a collection of objects. Suppose further that you become fixated and ask: which sack contains more items? To this end, you take an item from the first sack with your left hand; with your right hand, you take one from the second. It seems reasonable to think that, if the sacks are the same size, there will be no moment where one of your hands is empty. In other words, you will always be able to match an object from the first sack (left hand) with an object from the second sack (right hand).

In mathematics, this operation—consisting in pairing off the elements of two sets—is known as a bijection. This notion seems effective for deciding whether one sack contains more items than another. A sack containing all the letters of the alphabet, for instance, may be put in bijection with a sack containing the numbers from 1 to 26, but not with a sack containing all the words in the dictionary. Cantor's idea was to extend this definition to sets of infinite size. His question thus becomes: is it possible to put into bijection the sack containing all the natural numbers ($\mathbb{N}$) with the sack containing all the real numbers ($\mathbb{R}$)? It turns out the answer is negative: there are infinitely more real numbers than there are natural numbers. In mathematical jargon, we say that $\mathbb{N}$ has cardinality $\aleph_0$ whereas $\mathbb{R}$ has cardinality $\aleph_0^2$ (see Appendix 5).

A transfinite Turing machine, denoted $\mathbf{T}_{\omega_1}$, or more simply $\mathbf{T}_1$, is a generalisation of the classical Turing machine. Like the latter, it has a read-write head, an infinite tape, a set of states, and a transition function, but it differs in two crucial respects. First, the tape is of infinite cardinality, not merely unbounded in length. More precisely, we shall require that its ordinality remain less than ω_1, but below this bound, it may be extended as

often as necessary. Furthermore—and this is the most important point—$\mathbf{T}_1$ is capable of performing a number α of successive operations, provided that α remains strictly below the ordinal ω_1. Where a classical Turing machine would run indefinitely, $\mathbf{T}_1$ is able to reach the limit of the recursive process. This allows it, notably, to write each of the natural numbers one by one, or to halt after having reached the "end" of the tape (in a sense that will be clarified). Under the Continuum Hypothesis, $\mathbf{T}_1$ can also write an irrational number on the tape. However, it is unable to enumerate the entire set of real numbers, or even the set of points on a line segment, since these sets have cardinality $\aleph_0^2$.

Unlike machines endowed with an "Oracle" (see Section 5.3.4), $\mathbf{T}_1$ cannot, in a single unit of time, provide the answer to questions that would require performing an infinite number of iterations. Rather, it is we observers who, when contemplating a transfinite machine, are able to count the ω_1 operations needed for a recursion to be completed. In other words, the definition of a transfinite Turing machine does not so much alter the properties of the classical Turing machine in itself as it transforms our way of counting the potentially infinite time during which the machine executes. We shall return in more detail to this comparison with oracle-equipped Turing machines, but only after having formalised our definition of the transfinite Turing machine.

The extension of the Turing machine, as I develop it in the following pages, is closely related to the models of the Infinite Time Turing Machine (ITTM) and of Zeno Machines. At the end of this chapter, I shall set out the formal differences between my definition and those found in the existing literature.

10.2.1 Formal definition

Let us consider a tape composed of a countably infinite number of indexed cells. On the first of these appears the indelible symbol $\triangleright$. This symbol marks both the leftmost end of the tape and the starting point of the reading head. Each time $\mathbf{T}_1$ reads this symbol, the transition function instructs it with the command $\rightarrow$, ordering it to move one cell to the right, possibly (but not necessarily) with a reassignment of its current state. In some cases, the machine may loop indefinitely between the symbol $\triangleright$ and the second cell of the tape. For a transfinite Turing machine, this kind of phenomenon is not identical to an endless progression towards the right-hand side of the tape, as we shall soon observe.

Indeed, on the right side of the tape, after ω_0 "normal" cells, there lies a "special" cell displaying the indelible symbol $|_{\omega_0}\triangleright$ (which may be abbreviated as $|_0\triangleright$ when the context is sufficiently clear). A classical Turing machine would never reach such a symbol, but a transfinite machine may arrive at it after ω_0 operations. The state in which $\mathbf{T}_1$ finds itself upon reaching $|_0\triangleright$

depends on the infinitely many, countable states $q_1, q_2, \ldots$ it has traversed up to that point. If the sequence $q_1, q_2, \ldots$ converges to a state q, we consider the machine to be in state q upon reaching $|_0\triangleright$. Otherwise, the machine arrives at the symbol $|_0\triangleright$ in an "undefined state", denoted by ∞. Note that it does not matter whether the sequence $q_1, q_2, \ldots$ is considered from q_1 or from any q_n, since the convergence (or divergence) of $q_1, q_2, \ldots$ is equivalent to that of $q_n, q_{n+1}, \ldots$. The only condition is that the starting point of the sequence must occur at a time i posterior to the machine's last passage over the symbol $|_0\triangleright$ (if any). The ability of a transfinite machine to determine whether an infinite sequence of its states converges or not is perhaps its most important distinguishing feature compared to a classical Turing machine.

Machines stables et instables

Some functions f defined on the natural numbers are undefined for arguments tending towards infinity. A clear case is the function $f(n) = n \bmod 2$. Due to such singularities, one may expect that the output of a transfinite machine capable of computing functions may not always be well defined. As an example, consider a machine that writes the symbol "p" or "i" on the first cell of its tape, depending on whether the rest of the tape contains an even or, respectively, an odd number of symbols "a". On an input containing an infinite number of "a"'s, the result of the machine is undefined. As soon as ω_0 operations have been performed, the symbol "?" is written on the first cell of the tape. The transition function δ will then decide what to do when such a symbol is read. In general, a machine $\mathbf{T}_\alpha$ is said to be *unstable* if there exists at least one input x that causes one or more "?" symbols to appear on the tape, and *stable* otherwise. Likewise, a cell of the tape is said to be stable if the symbol "?" never appears there, and unstable if there exists at least one way for such a symbol to appear.

Once the reading head reaches the symbol $|_0\triangleright$, three scenarios may unfold: it may either enter a halting state, in which case the program terminates; or proceed to the right with the command $\rightarrow$, in which case the tape extends; or return to the left with the command $\leftarrow$. In the case where the instruction $\rightarrow$ is given, the tape extends by as many cells as necessary. The machine may then perform a finite number of operations, or reach the next symbol $|_0\triangleright$ via a further ω_0 iterations. This new symbol will behave exactly as the previous one. There is, moreover, no upper limit to the number of $|_0\triangleright$ symbols the machine may traverse, so long as they are countably infinite in number. Even when reaching the position $\omega_0\omega_0$, a machine $\mathbf{T}_1$ will encounter a new $|_0\triangleright$ symbol. Thus, there is no way for this machine to ever

reach the end of the tape on its right-hand side. Formally, we say the tape is bounded above by the limit ordinal ω_1, but that beneath this limit, it can grow arbitrarily large.

If the reading head encounters a $|_0\triangleright$ symbol and is given the instruction $\leftarrow$, it immediately returns to the symbol $\triangleright$. Note that from a $|_0\triangleright$ position, it would be impossible to perform "a single" move to the left, for in fact $\omega_0 - 1 = \omega_0$ (where "$-$" denotes left-subtraction). This is why the instruction $\leftarrow$ in this context necessarily amounts to a "return to the starting cell". This is the only situation in which an instantaneous jump may occur. In terms of time complexity (see section 3), we shall later see that a "return to the starting cell" may equally be interpreted as a countably infinite number of unit operations, as a single $\leftarrow$ operation, or as an instantaneous displacement — none of these interpretations having any effect on the overall complexity[7].

One might ask here why the $\leftarrow$ command, from a cell of type $|_0\triangleright$, performs a return to the starting cell, rather than a "jump" to the nearest $|_0\triangleright$ symbol. This choice is neither arbitrary nor crucial. On the one hand, it can be shown that the computational power afforded by a machine using this alternative definition would be equivalent to that provided by our own (in the precise sense that every computable function or semi-decidable set in one case is also so in the other). Moreover, both definitions would lead to equivalent time complexities, as shall later be demonstrated. Finally—and this is the decisive reason—the alternative definition proves more difficult to generalise to machines $\mathbf{T}_{\omega_2}, \mathbf{T}_{\omega_3}, \mathbf{T}_{\omega_{\omega_0}}, \ldots$ capable of executing more operations than $\mathbf{T}_{\omega_1}$.

If we consider $\mathbf{T}_{\omega_n}$, for instance (with n any ordinal), new symbols $|_1\triangleright, |_2\triangleright, \ldots, |_{n-1}\triangleright$ must be added to the tape, such that for every $0 \leq i < n$, the symbols $|_i\triangleright$ are spaced at distances of ω_i from one another. The rules governing these new symbols are identical to those previously described. The machine may thus arrive at them in a defined state or in state ∞, depending on whether the sequence of ω_i preceding states converges or diverges. When the $\leftarrow$ instruction is issued on any of these symbols, it allows a return to $\triangleright$, *and not to the previous symbol of type* $|_j\triangleright$ *to its left* (for any j). Indeed, there may not always exist a "previous" $|_j\triangleright$ symbol. (To see this, consider the particular case where $\omega_n = \omega_{\omega_0}$, which is a strongly inaccessible ordinal. When the reading head is positioned at the first $|_{\omega_0}\triangleright$ symbol corresponding to this ordinal, one would be hard-pressed to say what the closest $|_i\triangleright$ symbol to its left might be.)

Each time a $\leftarrow$ or $\rightarrow$ instruction is issued, the transition function may also assign a new state to the machine, though this is not mandatory. The state ∞, however, is a special state, which only appears when the sequence

[7] When defining time complexity for transfinite machines, we shall consider the size of operations in terms of cardinality. Sequences of operations of ordinality $\omega_0, \omega_0 + 1, \omega_0 2$ will all be assigned the same complexity $\aleph_0$.

of ω_0 states traversed by the machine is divergent.

All these remarks lead us to the formal definition below. For convenience, the notations $\mathbf{T}_{\omega_i}$ and $|_{\omega_i}\triangleright$ will sometimes be abbreviated as $\mathbf{T}_i$ and $|_i\triangleright$, where the context allows. For any ordinal α, we may also use the notations $\mathbf{T}_\alpha$ and $|_\alpha\triangleright$. Finally, let us denote by Σ^* the set of all (finite or infinite) strings constructed from an alphabet Σ, including the null string e of length zero.

Définition 11. *Let α be a limit ordinal. A transfinite Turing machine $\mathbf{T}_\alpha$ is a sextuple $(K, \Sigma, \delta, s, H, \alpha)$ where*

(a) K is a finite set of states, including the divergent state ∞;

(b) Σ is an alphabet containing the symbol $\sqcup$ (representing a blank cell), the symbols $\triangleright$, $|_\alpha\triangleright$ (marking the "limit" positions of the tape), the intermediate symbols $|_\beta\triangleright$ (defined for every limit ordinal $\beta \subset \alpha$), and the symbol "?" (signalling an indeterminacy), but excluding the symbols $\rightarrow$ and $\leftarrow$, which are used exclusively to represent head movements;

(c) $s \in K$ is the initial state of the machine;

(d) $H \subseteq K$ is the set of halting states;

(e) δ is a transition function from $K \setminus H \times \Sigma$ to $(K \setminus \{\infty\}) \times (\Sigma \cup \{\rightarrow, \leftarrow\})$ such that:

1. *the image of (q, σ) assigns a state q and an instruction σ of the type $\rightarrow$ (move right), $\leftarrow$ (move left), or $\sigma \in \Sigma$ (meaning the current cell must be overwritten with the symbol σ);*
2. *the image set of δ does not include the symbols "?", $\triangleright$, $|_\alpha\triangleright$ or $|_\beta\triangleright$ (for any $\beta \subset \alpha$);*
3. *for every $p \in K \setminus H$, $\delta(p, \triangleright) = (q, \rightarrow)$.*

For every limit ordinal $\beta \subset \alpha$, the following rules are also observed:

(i) when β successive $\rightarrow$ instructions are executed, then

- *(i.1) the reading head reaches the next symbol $|_\beta\triangleright$ on the tape (possibly passing over other symbols of the form $|_\gamma\triangleright$, where $\gamma \subset \beta$);*
- *(i.2) if the sequence of states $q_0, q_1, ...q_\gamma, ... \in K$ traversed to reach $|_\beta\triangleright$, starting from the last encountered $\triangleright$ or $|_\theta\triangleright$ symbol (for any θ), converges to a state $q \in K$, then the machine is in state q; otherwise, it is in state ∞;*

– *(i.3) if the sequence of symbols $\sigma_1, \sigma_2, ...$ successively written on a single cell during this process converges to a symbol $\sigma \in \Sigma$, then that symbol appears on the cell; otherwise, the symbol "?" is displayed;*

(ii) when a $\leftarrow$ instruction is executed from a $|_\beta\triangleright$ position on the tape, the reading head returns to the $\triangleright$ symbol;

(iii) the length of the tape and the total number of operations permitted by the machine are bounded above by α[8].

By convention, I have assumed that the set of states K is finite. It is important to note that a machine with an infinite set of states would also possess a transition function with an infinite number of possible inputs. This would correspond to an algorithm with an infinite number of lines of code. It is likely that the computational power of such a machine would exceed that of a Transfinite Turing Machine, provided that the size of K is at least as great as the ordinal α employed in the definition of the transfinite machine.

Définition 12. *Let $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ and let $x \in \Sigma^*$ be a string. We say that x is an input of $\mathbf{T}_\alpha$ if and only if its cardinal length is strictly less than α and it does not contain the symbols $\sqcup$, $\triangleright$, or $|_\beta\triangleright$ (defined for all limit ordinals $\beta \subseteq \alpha$). Moreover, the string present on the tape at the moment when $\mathbf{T}_\alpha$ halts (if it halts) is called the output.*

In the case where $\alpha = \omega_1$, the machine $\mathbf{T}_{\omega_1}$ is the one described at the beginning of this section. A machine $\mathbf{T}_{\omega_2}$, for its part, would possess a tape of uncountable cardinality. Provided the continuum hypothesis is accepted, such a tape could be modelled using a real line as index. To each real number r would correspond a position on the tape, which could either be blank $\sqcup$, or filled with a symbol from the alphabet.

At each time step, $\mathbf{T}_\alpha$ is in a state p and reads the letter σ_{actuel} corresponding to the current position of its reading head on the tape. If the transition function $\delta(p, \sigma_{\text{actuel}})$ yields $(q, \rightarrow)$, the reading head moves one cell to the right and the machine enters state q (in the particular case where $p = q$, the state remains unchanged). If the transition function yields $(q, \leftarrow)$, the machine enters state q and shifts one cell to the left, "returning" to the starting cell $\triangleright$ when this operation is executed from a location $\sigma_{\text{actuel}} = |_\theta\triangleright$ (for any θ). If the transition function returns (q, σ), the machine enters state q and overwrites the letter σ_{actuel} with the letter σ. In the special case where $\sigma_{\text{actuel}} = \sigma$, the tape remains unchanged. When $\sigma = \sqcup$, the machine erases the symbol read by the head and replaces it with the blank symbol $\sqcup$.

[8] The notation used in this definition is based on the definition of a Turing machine by Lewis R. H. and Papadimitriou C. H., *Elements of the Theory of Computation*, Prentice-Hall, Upper Saddle River, New Jersey, 1998, definition 4.1.1, p.181.

The execution of $\mathbf{T}_\alpha$ continues as long as the number β of iterations performed remains less than or equal to α. If, at any point, the machine enters a state in H, the process halts. If there exists no ordinal $\beta < \alpha$ such that this condition is met, the machine is said to loop "α-infinitely".

Définition 13. *Let $\gamma \subseteq \alpha$ be two limit ordinals, and let $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ be a transfinite Turing machine with input $x \in \Sigma*$. For each ordinal $\beta \subset \alpha$, denote by q_β the state of $\mathbf{T}_\alpha$ after β iterations. We say that $\mathbf{T}_\alpha$ loops γ-infinitely on x if there exists no ordinal $\beta \subset \gamma$ such that $q_\beta \in H$.*

Exemple 1. *Let us consider the Turing machine $\mathbf{T}_{\omega_1} = (K, \Sigma, \delta, s, \{h\}, \omega_1)$ where $K = \{q_0, q_1, h\}, \Sigma = \{\triangleright, |_0\triangleright, \sqcup, a, b\}, s = q_0$ and the transition function δ is defined by the following table.*

$q,$	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	$\|_0\triangleright$	$(q_0, \rightarrow)$
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	$\|_0\triangleright$	$(h, \rightarrow)$
∞	$\triangleright$	$(h, \rightarrow)$
∞	a	$(h, \rightarrow)$
∞	b	$(h, \rightarrow)$
∞	$\sqcup$	$(h, \rightarrow)$
∞	$\|_0\triangleright$	$(h, \rightarrow)$

Table 10.1

The program on the left scans the input from end to end, replacing the first blank symbol $\sqcup$ with a, the second blank with b, and so on, alternating cyclically.
Given a finite input $x \in \Sigma^*$ of length n, the machine will encounter an infinite sequence of blank symbols starting from position $n+1$. Upon reaching the symbol $|_0\triangleright$, the sequence of states traversed will form an endless alternation: $q_0, q_1, q_0, q_1, \ldots$ This sequence diverges, so the machine enters the state ∞. Since $\delta(\infty, |_0\triangleright) = (h, \rightarrow)$, and h is a halting state, this marks the termination of the process.
On the other hand, given an infinite-length input $x \in \Sigma^*$, the machine will encounter no blank symbols and will converge to state q_0 upon reaching $|_0\triangleright$. Now, we know that $\delta(q_0, |_0\triangleright) = (q_0, \rightarrow)$. The machine will thus continue scanning the tape until it finds a blank symbol $\sqcup$, surpassing as many occurrences of $|_0\triangleright$ as needed. Since the machine is of type T_{ω_1} and we have $|x| < \omega_1$, there will necessarily come a point at which infinitely many $\sqcup$ symbols appear. From that moment on, the machine will proceed exactly as in the finite input case and halt upon encountering the next $|_0\triangleright$.

Note that since δ is defined on the domain $K \setminus H \times \Sigma$, it must specify a value for every pair (q, σ). In practice, however, the pairs $(\infty, \triangleright)$, $(\infty, \sqcup)$, (∞, a), (∞, b) will never be encountered (see definition 11, point e.2) and may therefore be omitted from the table. I have included them here for the sake of completeness.

Exemple 2. *Let us now consider the Turing machine* $\mathbf{T}_{\omega_1} = (K, \Sigma, \delta, s, \{h\}, \omega_1)$ *where* $K = \{q_0, q_1, h\}, \Sigma = \{\triangleright, |_0\triangleright, \sqcup, a, b\}, s = q_0$ *and the transition function* δ *is given by the same table as in example 1, with the following exception:*

$q,$	σ	$\delta(q, \sigma)$
∞	$\|_0\triangleright$	$(q_1, \rightarrow)$

For any input $x \in \Sigma^*$ *of arbitrary length, the machine will behave just as in example 1, except that once it begins to write the alternating sequence* $a, b, a, b, \ldots$ *on the tape, it will never cease. Indeed,* $\mathbf{T}_{\omega_1}$ *will reach the next* $|_0\triangleright$ *cell in the state* ∞*. Since* $\delta(\infty, |_0\triangleright) = (q_1, \rightarrow)$*, the machine will continue moving to the right, endlessly inscribing* $a, b, a, b, \ldots$

10.2.2 Restrictions and Extensions of a Transfinite Machine

In what precise sense can one say that the transfinite machine is an extension of the classical Turing machine? In Example 1 of the previous section, the machine we constructed loops ω_0-infinitely for every possible input $x \in \Sigma^*$. In the second example, however, the machine loops ω_1-infinitely (and thus also ω_0-infinitely). Yet in both cases, the machines execute the "same" programme:

```
programme1(x)
Move to the first blank cell
while(true)
        write "a"
        write "b"
end
```

Below $\tau < \omega_0$ steps, the transfinite machines implementing the above programme behave in a perfectly indistinguishable manner, but differences begin to emerge after ω_0 units of time. The machine from Example 1 halts, whereas the second continues endlessly writing a and b. This shows that the finite programmes we commonly use, if executed by a transfinite machine, would be ambiguous. One would need, each time, to specify their behaviour at the limit points of the tape (those containing the symbols $|_\theta\triangleright$ for all limit ordinals $\theta \subset \alpha$).

Définition 14. *Let* α *and* β *be two limit ordinals such that* $\alpha \subset \beta$*. We say that* $\mathbf{T}_\beta = (K', \Sigma', \delta', s, H, \beta)$ *is a* β*-extension of* $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$*, and conversely that* $\mathbf{T}_\alpha$ *is an* α*-reduction of* $\mathbf{T}_\beta$*, if and only if:*

(a) $K' = K \cup \{\infty\}$*;*

(b) $\Sigma' = \Sigma \cup \{|_\gamma \triangleright\}$ *for every limit ordinal* $\beta \subseteq \gamma \subseteq \alpha$;

(c) *the function* δ' *is defined such that for every* (q, σ) *in the domain of* δ, *we have* $\delta'(q, \sigma) = \delta(q, \sigma)$;

Note that, for a fixed β, the machine $\mathbf{T}_\beta$ is the β-extension of only one machine $\mathbf{T}_\alpha$: namely, the one obtained from $\mathbf{T}_\beta$ by removing every reference to symbols $|_\gamma \triangleright$ of order greater than α. However, the converse does not hold. The machine $\mathbf{T}_\alpha$ has a multiplicity of β-extensions. This asymmetry suggests that it is possible to determine the behaviour of a reduced machine based on the one from which it is derived—but not the behaviour of an extended machine from the one it extends.

Définition 15. *Let* $\tau \subset \alpha$ *be two ordinals and* $x \in \Sigma^*$ *a string of length* β. *The truncation at* τ *of* x *is the string* x *obtained by removing all characters of* x' *beyond the position* τ.

Définition 16. *Let* $\mathbf{T}_\alpha$ *be a machine. The configuration of* $\mathbf{T}_\alpha$ *at a given moment is the 3-tuple* (p, x_1, x_2), *where* $p \in K$ *represents the current state of the machine,* $x_1 \in \Sigma^* \cup \{\triangleright, |_0\triangleright, |_1\triangleright, ...\}$ *is the string of characters to the left of the reading head (including the cell on which it rests), and* $x_2 \in \Sigma^* \cup \{|_0\triangleright, |_1\triangleright, ...\}$ *is the string of characters to the right of the reading head. By convention, a single* $\sqcup$ *is written once the tape consists only of white cells.*

Exemple 3. *Let us revisit the machine* $\mathbf{T}_{\omega_1}$ *from example 1 and examine its configuration after four operations on an empty input. It starts at* $\triangleright$, *reads a symbol* $\sqcup$, *then successively writes a and b. It is now on a white cell in state* q_1. *The configuration is thus described by* $(q_1, \triangleright \sqcup\, a\, b \sqcup, \sqcup)$.

Proposition 5. *Let* $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ *be a machine and* $\mathbf{T}_\beta = (K', \Sigma', \delta', s, H, \beta)$ *a* β*-extension. Suppose that* $\mathbf{T}_\beta$ *is run on the input* $x \in \Sigma^*$ *and* $\mathbf{T}_\alpha$ *is run on the truncation at* τ *of* x. *After* τ *operations on each side, the two machines will be in the same configurations.*

Proof. At the initial time, $\mathbf{T}_\alpha$ and $\mathbf{T}_\beta$ both start with the same initial state s and the same first τ symbols on the tape (by definition 15). Their transition functions δ and δ' yield the same output for every $p \in K \cap K'$ and $\sigma \in \Sigma \cap \Sigma'$ (definition 14.c). Therefore, the only symbols where differences might arise are the symbols $|_\alpha \triangleright \in \Sigma'$, any symbols $|_\gamma \triangleright \in \Sigma'$ (for any γ such that $\alpha \subset \gamma \subset \beta$), and the symbols of x beyond the τ-th symbol. However, since $\tau \subset \alpha$, the machine $\mathbf{T}_\beta$ has not had time to reach any of these symbols.

Note that if $\alpha = \omega_0$, the functions δ and δ' may also differ on $\in K' \setminus K = \{\infty\}$ (definition 14.a). However, the symbol ∞ will only come into play when $\mathbf{T}_\beta$ has reached the symbol $|_0\triangleright$, which does not happen within $\tau \subset \alpha = \omega_0$ operations. □

Corollary 1. *Let γ, α, β be three limit ordinals such that $\gamma \subseteq \alpha \subset \beta$. The machine $\mathbf{T}_\alpha$ loops γ-infinitely on an input $x \in \Sigma^*$ if and only if its β-extension $\mathbf{T}_\beta$ loops γ-infinitely on x.*

Proof. First, note that x has strictly smaller ordinality than α, otherwise it would not be a valid input for $\mathbf{T}_\alpha$. For any $\tau \subset \gamma$, $\mathbf{T}_\alpha$ does not halt after τ operations (by definition 13 of γ-looping). Now, by the previous proposition, $\mathbf{T}_\alpha$ and $\mathbf{T}_\beta$ have the same configuration after τ operations. Thus, for any $\tau \subset \gamma$, we conclude that $\mathbf{T}_\alpha$ does not halt after τ operations either. The converse follows similarly. □

The above proposition guarantees that expressions like "the machine $\mathbf{T}_\alpha$ loops γ-infinitely on x" refer not only to a property of $\mathbf{T}_\alpha$, but more generally, to a property of the algorithm implemented by this machine. In other words, we can confidently say that if a certain algorithm loops γ-infinitely on a given input, then the algorithm retains this characteristic regardless of the size of the tape and the number of iterations allotted to the machines implementing it.

One question that might arise, however, is the following: Are there properties that are preserved regardless of the size of the input x? In particular, what happens if an algorithm capable of solving the 3-SAT problem in finite time on a classical Turing machine is executed by a machine $\mathbf{T}_1$ on a countably infinite input? Will it halt in countably infinite time or will it loop ω_1-infinitely? Will we obtain a solution to 3-SAT$_{\aleph_0}$? Answering such questions is far from trivial, as increasing the size of the input may also modify the semantics of the program. To highlight this limitation, we will derive undecidability results for transfinite machines, analogous to the halting problem and Rice's theorem. But before that, it is necessary to ask what the computational power of the transfinite machines we have constructed so far is. What types of languages are they capable of deciding or semi-deciding? Can they be compared to oracle Turing machines? And based on this, is it possible to construct a universal transfinite Turing machine?

10.2.3 Computational Power

Any function that is computable by a classical Turing machine is also computable by a transfinite Turing machine. This fact is established by proposition 5. The same holds for all decidable or semi-decidable languages. In fact, differences are likely to emerge only once at least ω_0 operations are performed. It is known that these differences do indeed appear, due to the existence of indeterminate functions (see the box 10.2.1 on unstable machines). As a result, one may expect that a subset of the languages $\mathcal{L}_\alpha$, consisting of words of lengths less than or equal to α, are "non-computable" (in a sense perhaps different from the halting problem). Formally, we write

$\mathcal{L}_0 \subseteq \mathcal{D}_\alpha \subseteq \mathcal{L}_\alpha$, where $\mathcal{D}_\alpha$ denotes the set of languages, composed of words of size less than or equal to α, which are decidable or semi-decidable by a machine $\mathbf{T}_\alpha$.

In example 2 from the first section, we constructed a transfinite machine that successively writes the characters a and b on the tape, until it reaches the symbol $|_0\triangleright$. At each step, this machine moves from state q_1 to q_2, and this indefinitely. When it reaches the first symbol $|_0\triangleright$, the machine enters the indeterminate state ∞. However, one might remark that the sequence of states $q_1, q_2, q_1, q_2, ...$, though divergent, oscillates in a regular manner. Rather than entering state ∞, the machine could enter state $\infty_{1,2}$. One would then speak of a k-cyclic machine, where k is the maximum size of the cycles it would be capable of identifying.

Définition 17. *Let α be a limit ordinal and $k \in \mathbb{N}$. A k-cyclic Turing machine $\mathbf{C}_\alpha = (K', \Sigma', \delta', s, H, \alpha)$ is a transfinite Turing machine $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ to which the following additions (or modifications) have been made:*

- *K, Σ, s, and H are defined in the same way as for a Turing machine $\mathbf{T}_\alpha$ (see definition 11).*
- *For every subset of states $\{q_1, q_2, ..., q_n\} \in \mathbb{P}(K)$ (with $2 \leq n \leq k$), there exists in K' a cyclic state $\infty_{1,2,...,n}$ in K'.*
- *For every subset of characters $\{q_1, q_2, ..., q_n\} \in \mathbb{P}(\Sigma)$ (with $2 \leq n \leq k$), there exists in Σ' a cyclic character $?_{1,2,...,n}$ in Σ'.*
- *(e) The transition function δ is a function from $K' \setminus \{\infty\} \times \Sigma'$ to $K' \times \Sigma \cap \{\rightarrow, \leftarrow\}$. There is no image (σ, q) of the function δ' such that q is a cyclic state or σ is a cyclic character. Rules e.1 to e.3 are preserved.*
- *(i.2) If the sequence of states $q_0, q_1, ...q_\gamma, ... \in K$ traversed to reach $|_\beta\triangleright$, from the last symbol $\triangleright$ or $|_\theta\triangleright$ encountered (with θ arbitrary), converges to a cycle of states $q_1, q_2, ..., q_n \in K$ (with $2 \leq n \leq k$), then the machine is in the cyclic state $\infty_{1,2,...,n}$, otherwise it is in the state ∞;*
- *(i.3) If the sequence of symbols $\sigma_1, \sigma_2, ...$ succeeding each other on the same cell during this process converges to a cycle of symbols $\sigma_1, \sigma_2, ..., \sigma_n \in \Sigma$ (with $2 \leq n \leq k$), then the cell presents the cyclic symbol $?_{\sigma_1,\sigma_2,...,\sigma_n}$, otherwise it shows the symbol "?";*

Contrary to appearances, the (semi)-decidable languages by a k-cyclic machine are identical to the (semi)-decidable languages by a transfinite Turing machine. The reason is that a non-cyclic Turing machine can write on

its tape m-tuples $(q_1, q_2, ..., q_m)$ of sizes $2 \leq m \leq k$. At any point, these m-tuples can be constructed in such a way as to "remember" the m last states traversed by the Turing machine. If, after an infinite number of operations, all these m-tuples are replaced by the symbol "?", this means that the successive states $q_1, q_2, ...$ of $\mathbf{C}_\alpha$ do not converge to a cycle of length m. If there is at least one m-tuple that stabilises, then we know that the successive states of $\mathbf{C}_\alpha$ converge to a cycle whose length is a divisor of m. By exploiting this information, the non-cyclic machine can thus reproduce the behaviour of a cyclic machine. Of course, some technical issues need to be addressed, such as: how should these m-tuples be constructed? Where should they be stored? How should they be used? The following theorem provides the details of this approach.

Proposition 6. *Let $\mathbf{C}_\alpha = (K', \Sigma', \delta', s', H', \alpha)$ be a k-cyclic machine, with $k \in \mathbb{N}$. There exists a Turing machine $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ with a single tape that simulates $\mathbf{C}_\alpha$. More precisely, suppose that $\mathbf{C}_\alpha$ starts with $x \in \Sigma^*$ on its tape and ends with $y \in \Sigma^*$. By giving x as input to $\mathbf{T}_\alpha$, the latter will also finish with y as output. Furthermore, the execution times of both machines are equipotent.*

Proof. We begin by proving the case $\alpha = \omega_0$, after which we will proceed by induction. Without loss of generality, we can consider that $\mathbf{T}_{\omega_1}$ is in normal form, so that the cells located before the first symbol $|_0\triangleright$ are empty. Let C_{ω_0+1} be the first of these blank cells. Similarly, let C_i denote the i-th cell directly after the symbol $\triangleright$. The machine $\mathbf{T}_\alpha$ has an alphabet $\Sigma = \Sigma' \cup \Sigma'^2 \cup \cdots \cup \Sigma'^k \cup \{1, 2, ..., k\}$, which allows it to write m-tuples of the form $(\sigma_1, \sigma_2, ..., \sigma_m)$, with $\sigma_1, \sigma_2, ... \in \Sigma'$ and $2 \leq m \leq k$. Our machine is also capable of writing digits on the tape, ranging from 1 to k, which we will refer to as "cyclicity indices." If one or more digits between 1 and k are already included in Σ', we choose an encoding for these digits (we will assume here, without loss of generality, that this is not the case). For every state $q \in K$, we also assume there exists a unique encoding q^* made with the letters of Σ'. For each cyclic state $\infty_\mu \in K'$ (with $\mu = (q_1, q_2, ..., q_m)$), we create a state q_μ in K such that $\delta'(\infty_\mu, \sigma) = \delta(q_\mu, \sigma)$ for all $\sigma \in \Sigma'$. We also create a state $q_\infty \in K$ such that $\delta'(\infty, \sigma) = \delta(q_\infty, \sigma)$ for all $\sigma \in \Sigma'$. Finally, we add symbols $?_\mu$ to K' such that, for every state $q \in K'$ and for every m-tuple $\mu = (\sigma_1, \sigma_2, ..., \sigma_m)$, we have $\delta'(q, ?_\mu) = \delta(q, ?_\mu)$.

Before starting the simulation, $\mathbf{T}_\alpha$ presents a blank tape. On cell C_1, it writes the representation s'^* of the state s' in which the machine $\mathbf{C}_\alpha$ initially resides. On cells C_2 and C_3, it writes the number 2 and the 2-tuple (s'^*, s'^*). On cells C_4 and C_5, it writes the number 3 and the 3-tuple (s'^*, s'^*, s'^*), and so on... When it reaches cells C_{2k-2} and C_{2k-1}, it writes the number k and the k-tuple $(s'^*, s'^*, ..., s'^*)$. On the cells beyond C_{2k}, $\mathbf{T}_\alpha$ will represent the input x, following an analogous construction.

Let $\theta \subseteq \omega_0$ be the size of x, and $(x_i)_{i<\theta}$ the sequence of characters making up x. For every $1 \leq n \leq \theta$, the machine writes x_n on cell C_{2nk}, then the digit 2 on cell C_{2nk+1}, the 2-tuple (x_n, x_n) on cell C_{2nk+2}, followed by the number 3 and the 3-tuple (x_n, x_n, x_n), etc. It proceeds in this way until it has written the number k and the k-tuple $(x_n, x_n, ..., x_n)$ on cells $C_{(n+1)k-2}$ and $C_{(n+1)k-1}$. The attached diagram represents the state of the tape at the initial state.

Our machine $\mathbf{T}_\alpha$ will then enter state s', return to the beginning of the tape, and simulate the behaviour of $\mathbf{C}_\alpha$ in the following manner:

(a) Every time $\mathbf{C}_\alpha$ reads the character written on cell C_n, $\mathbf{T}_\alpha$ reads the character written on cell C_{2kn}.

(b) Every time $\mathbf{C}_\alpha$ executes the operation $\rightarrow$, $\mathbf{T}_\alpha$ makes $2k$ moves to the right.

(c) Every time $\mathbf{C}_\alpha$ executes the operation $\leftarrow$, $\mathbf{T}_\alpha$ makes $2k$ moves to the left.

(d) Every time $\mathbf{C}_\alpha$ writes a symbol $\sigma \in \Sigma'$ on cell C_n, $\mathbf{T}_\alpha$ writes the symbol σ on cell C_{2nk}. Furthermore, it edits the $2k$ cells following C_n according to the rules below. For all $1 \leq m < k$:

- $\mathbf{T}_\alpha$ reads the cyclicity index i present on cell $C_{2nk+2m-2}$ and replaces it by $i+1$. If $i+1 > m$, it replaces it with a 1.
- $\mathbf{T}_\alpha$ goes to cell $C_{2nk+2m-1}$ and replaces the m-tuple $(\sigma_1, \sigma_2, ..., \sigma_m)$ present on this cell by an identical m-tuple, but where σ_i is replaced by σ.
- $\mathbf{T}_\alpha$ returns to cell C_{2nk} in the same state it was in when it arrived there.

(e) Every time $\mathbf{C}_\alpha$ moves from state p to state q (whether distinct or not), $\mathbf{T}_\alpha$ enters state q. Furthermore, it will edit the first $2k$ cells of the tape according to the rules below. For all $0 \leq m < k-1$:

- $\mathbf{T}_\alpha$ reads the cyclicity index i present on cell C_m and replaces it with $i+1$. If $i+1 > m$, it replaces it with a 1.

- $\mathbf{T}_\alpha$ goes to cell C_{m+1} and replaces the m-tuple $(q_1^*, q_2^*, ..., q_m^*)$ present on this cell by an identical m-tuple, but where q_i^* is replaced by q^*.
- $\mathbf{T}_\alpha$ then returns to state q and goes to the cell where it was when it entered state q.

(f) If $\mathbf{C}_\alpha$ enters a terminal state $h \in H'$, then $\mathbf{T}_\alpha$ will also enter a terminal state. But before doing so, it will restructure the tape. The symbol $\sigma \in \Sigma$ written on cell C_{2nk} will be transposed onto cell C_n, for all $n \subseteq \theta$.

(g) Every time $\mathbf{T}_\alpha$ reaches the symbol $|_0\triangleright$ in state ∞, it does the following:

1. $\mathbf{T}_\alpha$ moves to the symbol $\triangleright$ and successively reads the cells C_{2nk} (for $1 \leq n \leq \theta$). If the cell C_{2nk} contains the symbol "?", $\mathbf{T}_\alpha$ looks at the cells C_{2nk+1} to $C_{2(n+1)k-1}$ one by one (in ascending order). If one of these does not contain the symbol "?", but an m-tuple $(\sigma_1, \sigma_2, ..., \sigma_m)$, $\mathbf{T}_\alpha$ writes the symbol $?_{\sigma_1,\sigma_2,...,\sigma_m}$ in the cell C_{2nk}. Otherwise, $\mathbf{T}_\alpha$ leaves "?" in this cell.
2. $\mathbf{T}_\alpha$ then moves to cell C_1. If this cell contains the symbol "q^*", $\mathbf{T}_\alpha$ transitions to state q. Otherwise, it reads the cells C_2 to C_{2kn} one by one (and in ascending order). The first cell encountered that does not contain the symbol "?" causes $\mathbf{T}_\alpha$ to transition to state q_μ, where μ is the m-tuple written in that cell. If all cells from C_2 to C_{2kn} contain the symbol "?", then $\mathbf{T}_\alpha$ enters state q_∞.
3. $\mathbf{T}_\alpha$ performs the operation $\rightarrow$ infinitely many times, never changing state. It thus reaches the symbol $|_0\triangleright$ in the state it had at the end of the previous step. If this state is q_∞, then $\mathbf{T}_\alpha$ behaves in the same way as $\mathbf{C}_\alpha$ would have in state ∞. If the state is of the type q_μ, where μ is an m-tuple (with $1 < m < k$), then $\mathbf{T}_\alpha$ behaves in the same way as $\mathbf{C}_\alpha$ would have in state ∞_μ.

We do not give the exact construction of the transition function for $\mathbf{T}_\alpha$ associated with these rules, but we know that such a function exists. Indeed, all these rules are defined recursively. They can thus be implemented by a suitably chosen classical Turing machine $\mathbf{T_0}$. Now, by Proposition 5, $\mathbf{T_0}$ behaves like $\mathbf{T}_\alpha$ for a finite number of steps. We conclude that $\mathbf{T}_\alpha$ can implement these rules for an arbitrarily large finite number of steps. In the case where $\mathbf{C}_\alpha$ reaches the symbol $|_0\triangleright$, the only rule affected is rule (g). This rule asks for a return to $\triangleright$, which can be done by a transfinite machine using the operation $\leftarrow$. The other operations prescribed by rule (g) are defined recursively, and thus executable by $\mathbf{T}_\alpha$ for the same reasons mentioned

above. We therefore conclude that these rules are all implementable on a transfinite Turing machine.

It is easy to see that if $\mathbf{C}_\alpha$ halts on y in a finite number of steps, then the machine $\mathbf{T}_\alpha$ also halts on y. Indeed, rule (f) asks to retain only the contents of cells C_{2kn}. Now, by rules (a) to (d), these cells always contain the same symbols as the cells C_n of $\mathbf{C}_\alpha$. What remains to be shown is that $\mathbf{T}_\alpha$ can also simulate the behaviour of $\mathbf{C}_\alpha$ when the latter performs an infinite number of operations. The issue is that when $\mathbf{C}_\alpha$ reaches the symbol $|_0\triangleright$ in a given state q (whether cyclic or not), $\mathbf{T}_\alpha$, for its part, reaches the symbol $|_0\triangleright$ in state ∞. Indeed, rules (d) and (e), in order to be executed, require $\mathbf{T}_\alpha$ to pass through various intermediate states, so that the sequence of states traversed to reach $|_0\triangleright$ no longer converges. To address this difficulty, the cell C_1 of $\mathbf{T}_\alpha$ is constructed to represent the state of $\mathbf{C}_\alpha$ at any point in the process (by virtue of rule (e)). If the states $q_1, q_2, ...$ traversed by $\mathbf{C}_\alpha$ converge to q, then the symbols $q_1^*, q_2^*, ...$ in cell C_1 will stabilise towards the representation q^* of q. By reading this cell, $\mathbf{T}_\alpha$ can thus retrieve the information that the state of $\mathbf{C}_\alpha$ does indeed converge to q. This is precisely the intent of rules $(g.2)$ and $(g.3)$.

Let us note that if C_1 contains the symbol "?", this can only mean two things: (i) either $\mathbf{C}_\alpha$ has reached $|_0\triangleright$ in state ∞, (ii) or $\mathbf{C}_\alpha$ has reached it in a cyclic state ∞_* of length k. In case (i), the states $q_1, q_2, ...$ traversed by $\mathbf{C}_\alpha$ do not converge to any cycle of length k, so none of the m-tuples $(q_1^*, q_2^*, ...)$ obtained with rule (e) converge to a stable m-tuple. Consequently, the cells from C_2 to C_{2kn-1} (inclusive) all contain the indeterminate symbol "?". In case (ii), however, at least one of the m-tuples on cells C_2 to C_{2kn} must have stabilised. It is then enough to find the smallest of these to know the cycle of states to which $\mathbf{C}$ has converged, which happens necessarily if we traverse cells C_2 to C_{2k} in ascending order.

A final guarantee that must be provided is that the cells C_{2nk} of $\mathbf{T}_\alpha$ remain identical to the cells C_n of $\mathbf{C}_\alpha$ after an infinite number of steps. Indeed, when $\mathbf{C}_\alpha$ reaches $|_0\triangleright$, some cells of its tape may contain a cyclic symbol $?_\mu$ (where μ denotes an m-tuple of letters from Σ'). Since $\mathbf{T}_\alpha$ is not a cyclic machine, the corresponding cells C_{2nk} on its tape will contain, not the symbol $?_\mu$ (which does not exist for $\mathbf{T}_\alpha$), but the symbol "?". Therefore, behavioural differences between the two machines may appear. To mitigate this problem, rule $(e.2)$ asks to write, for every $m \leq k$, an m-tuple $(\sigma_1, \sigma_2, ..., \sigma_n)$ capturing the values $\sigma_1, \sigma_2, ..., \sigma_n \in \Sigma'$ taken by cell C_{2nk} over time. If these values converge to a cycle of length $l \leq k$, at least one of the m-tuples must stabilise. By reading this m-tuple, rule $(g.1)$ then allows us to recover the symbol $?_\mu$ to be written on cell C_{2kn}.

All these observations allow us to prove that $\mathbf{T}_\alpha$ will halt on y whenever $\mathbf{C}_\alpha$ halts on y. To check the equivalence of execution times, it must be pointed out that, if $\mathbf{C}_\alpha$ takes an infinite countable amount of time, $\mathbf{T}_\alpha$ will also take an infinite countable amount of time. Indeed, every time $\mathbf{C}_\alpha$

passes through $|_0\triangleright$, machine $\mathbf{T}_\alpha$ will pass at most twice through $|_0\triangleright_\alpha$ (and at least once). In the special case where $\mathbf{C}_\alpha$ halts on y in finite time, this time is not equivalent to the time taken by $\mathbf{T}_\alpha$ (because the latter performs additional operations due to rules (b) to (f)). However, one could construct another machine $\mathbf{T'}_\alpha$ that halts on y in exactly the same time. This machine would be defined using the same sextuple $(K', \Sigma', \delta', s', H', \alpha)$ as machine $\mathbf{C}_\alpha$ (since, for finite times, $\mathbf{C}_\alpha$ behaves like a non-cyclic machine).

All of this demonstrates that $\mathbf{C}_\alpha$ can be simulated by $\mathbf{T}_\alpha$ in the case $\alpha = \omega_0$ in an equivalent time. For the cases $\alpha > \omega_0$, we proceed by induction. Suppose it is true for a certain cardinal $\aleph_n$. We will prove that it is also true for the cardinal $\aleph_{n+1}$ (within the framework of this proof, we consider cardinals as limit ordinals). To do so, we first assume, without loss of generality, that $\mathbf{C}_\alpha$ is in normal form. Therefore, there are only blank cells beyond the symbol $|_{\aleph_n}\triangleright$. If $\mathbf{C}_\alpha$ takes fewer than $\aleph_n$ steps, the induction hypothesis allows us to construct a machine $\mathbf{T}_\alpha$ that simulates $\mathbf{C}_\alpha$. Otherwise, we can construct a machine $\mathbf{T}_\alpha$ by following the same rules (a) to (g) described above, but using $|_{\aleph_n}\triangleright$ instead of $|_0\triangleright$. We do not write the details, but it is easy to see that this approach proves the property for all $\alpha \subset \omega_0^{\omega_0}$. In the special case where α is an inaccessible ordinal, such as $\omega_0^{\omega_0}$, the classical induction proof turns out to be technically insufficient to draw conclusions. □

10.2.4 Minimal Representation

The definition of a machine $\mathbf{T}_1$ given by Definition 11 stipulates that the tape may extend indefinitely, as long as its length does not exceed the ordinal ω_1. Now, it is known that an infinite set can be placed in a one-to-one correspondence with any of its subsets. Therefore, it is theoretically possible to compress all the information on the tape between the symbol $\triangleright$ and the first occurrence of $|_0\triangleright$. Suppose, for instance, that the tape contains a sequence of $\omega_0^{\omega_0}$ characters. This sequence can be viewed as ω_0 ordered sequences of length ω_0, denoted by $(a_{i,j})_{i\in\mathbb{N},j\in\mathbb{N}}$ (with $a_{i,j} \in \Sigma$). Explicitly, we have $a_{0,0}$, $a_{0,1}$, $a_{0,2}$, ..., $a_{1,0}$, $a_{1,1}$, $a_{1,2}$, ..., $a_{n,0}$, $a_{n,1}$, $a_{n,2}$, ... To compress this information, it suffices to place each $a_{i,j}$ in the cell given by $\pi^2(i,j) = (i+j+1)(i+j)/2 + j$, Cantor's pairing function. This approach is generalisable to any type of tape, except those bounded by an inaccessible ordinal (such as ω_0 or $\omega_0^{\omega_0}$). In such cases, no ordinal $\beta \subset \alpha$ of the same cardinality as α exists with which to construct a bijection. The tape of a classical Turing machine $\mathbf{T}_0$, in particular, cannot be put in one-to-one correspondence with a tape of finite size.

Proposition 7. *Let α be a limit ordinal. If α is not inaccessible (weakly or strongly), let β be the largest limit ordinal such that $\beta \subset \alpha$. Any decided or semi-decided language and any function computed by a machine $\mathbf{T}_\alpha$ can*

be decided, semi-decided, or computed (respectively) by a machine $\mathbf{T}_\alpha$ *that would use only the first* β *cells of its tape.*

Proof. The proof that a machine T using its entire tape can be simulated by a machine T', which only uses the first β cells of its tape, follows from the discussion above.

Since it is possible to compress all of the information from the original tape onto a tape of length β, it is clear that it is not strictly necessary to use any additional cells. It suffices to divide the tape of length β in machine T' into sub-tapes $R_1, R_2, R_3, \ldots$. The sub-tape R_1 of T' encodes those cells of T's tape that are located between the symbol $\triangleright$ and the symbol $|_\beta\triangleright$; the sub-tape R_2 contains the cells between the symbols $|_\beta\triangleright$ and $|_{\beta 2}\triangleright$, and so on.

At the beginning of execution, machine T' is positioned on the leftmost cell of the tape, marked $\triangleright$. When the operations $\rightarrow$ or $\leftarrow$ are performed, it moves to the next cell to the right or to the left within the sub-tape R_1, skipping all cells that belong to other sub-tapes.

When T' reaches the symbol $|_\beta\triangleright$ at the end of the first β cells, it must be brought back to the initial position. It is necessary to remember that it has reached the symbol $|_\beta\triangleright$, because from this point onward, each time the operations $\rightarrow$ or $\leftarrow$ are executed, the read head moves to the next cell to the right or to the left within sub-tape R_2, skipping all cells belonging to other sub-tapes. If, at any point, the read head is on the initial cell $\triangleright$ and the $\leftarrow$ operation is executed, the machine remains on the $\triangleright$ symbol but memorises that it has returned to sub-tape R_1, rather than R_2.

To keep track of which sub-tape the machine should operate on, the simplest solution is to duplicate each of the states $q_1, q_2, \ldots$ of the original machine T for each existing sub-tape. The machine T' will then be in state $q_{i,j}$ if the simulated machine T would be in state q_i and the read head is on the portion of the tape, as represented in T', corresponding to the j-th sub-tape. □

Normal Form

A machine $\mathbf{T}_\alpha$ is said to be in *normal form* if, for every cardinal $\beta \subset \alpha$, the cells between the first instance of $|_\beta\triangleright$ and the first instance of $|_{\beta+1}\triangleright$ are blank at the beginning of the execution of $\mathbf{T}_\alpha$ and remain so throughout the process. Proposition 8 guarantees that any decided or semi-decided language and any function computed by a machine $\mathbf{T}_\alpha$ can be decided, semi-decided, or computed (respectively) by a machine $\mathbf{T'}_\alpha$ in normal form.

Définition 18. *Let $k \in \mathbb{N}$ and β be a limit ordinal. A minimal Turing machine $\mathbf{M}\beta = (K', \Sigma, \delta', s, H, \beta)$ is a transfinite Turing machine $\mathbf{T}\beta = (K, \Sigma, \delta, s, H, \beta)$ in normal form to which the following additions (or modifications) have been made:*

- *K, Σ, s, and H are defined as for a transfinite Turing machine (see Definition 11).*
- *K' contains all the elements of K as well as a symbol $\mathbf{1}$ representing the convergent state;*
- *(e) The transition function δ' is a function from $K' \setminus \infty, \mathbf{1} \times \Sigma'$ to $K' \times \Sigma \cap \rightarrow, \leftarrow$. Rules e.1 to e.3 are retained.*
- *(i.2) If the sequence of states $q_0, q_1, \ldots, q_\gamma, \ldots \in K'$ traversed to reach $|\beta\triangleright$, starting from the last encountered symbol $\triangleright$ or $|\theta\triangleright$ (for any θ), converges to a state $q \in K'$, then the machine is in the convergent state $\mathbf{1}$; otherwise, it is in the state ∞.*
- *After reaching the symbol $|\beta\triangleright$, the machine returns to the initial position $\triangleright$.*

Proposition 8. *Let $\mathbf{T}\alpha = (K, \Sigma, \delta, s, H, \beta)$ be a Turing machine. There exists a minimal machine $\mathbf{M}\beta = (K', \Sigma', \delta', s', H', \beta)$ that simulates $\mathbf{T}\alpha$. More precisely, suppose that $\mathbf{M}\beta$ begins with $x \in \Sigma$ on its tape and halts with $y \in \Sigma$. Given x as input to $\mathbf{T}\alpha$, the latter also terminates with y as output. Moreover, the running times of the two machines are equipotent.*

Proof. In Proposition 7, I showed that a machine $\mathbf{T}\alpha$ could be simulated by a Turing machine $\mathbf{T}\beta$. It then suffices to observe that this machine $\mathbf{T}\beta$ can be written in minimal form. □

From a practical perspective, note that a machine in minimal representation must always retain at least one blank cell $\sqcup$. In the absence of such a cell, its tape is saturated, making it impossible to add new symbols without deleting existing ones. Indeed, there is no spare space in which to store the value of a symbol being overwritten by another, and this information is irretrievably lost (unless the erased symbol can be recovered by performing a computation on the remaining symbols or by encoding it in a special machine state—though these are special cases).

However, as long as there is at least one blank cell $\sqcup$, the machine can shift all symbols one position to the right. To see this, suppose without loss of generality that the tape is of the form $\triangleright\sqcup a_1, a_2, a_3, \ldots$, with $\sqcup$ in the first position (an entirely analogous argument applies if this is not the case) and $a_i \in \Sigma$. The machine begins by copying the symbol a_1 into the blank space, yielding $\triangleright a_1 a_1, a_2, a_3, \ldots$. It then erases the second a_1.

10.2.5 The k-tape Machine

Using reasoning very similar to that used for minimal machines, one can show that an extended machine $\mathbf{T}\alpha^k$ with k tapes and k read heads possesses no greater computational power than a single-tape machine. To show this, let us first specify how $\mathbf{T}\alpha^k$ operates.

Définition 19. *Let α be a limit ordinal and $k \in \mathbb{N}$. A machine $\mathbf{T}\alpha^k$ is a sextuple $(K, \Sigma, \delta, s, H, \alpha)$, where K, Σ, s and H are defined as in the case of a transfinite Turing machine $\mathbf{T}\alpha$ (see Definition 11). The transition function δ is a function from $K \setminus H \times \Sigma^k$ to $K \setminus \infty \times (\Sigma \cap \rightarrow, \leftarrow)^k$. For all states $p \in K$ and all k-tuples $(\sigma_1, \sigma_2, ..., \sigma_k) \in \Sigma^k$, $\delta(p, (\sigma_1, \sigma_2, ..., \sigma_k)) = \delta(q, (\sigma'_1, \sigma'_2, ..., \sigma'_k))$ means that the machine transitions from state q to p, and the i-th read head replaces the symbol σ_i at its current position by the symbol σ'_i. Moreover, δ follows rules (i), (ii) and (iii) of Definition 11, as well as the specific conditions regarding special symbols (11.e.1, 11.e.2, and 11.e.3).*

Thus, the machine has a single global state register for all operations, but its read heads can each perform an independent action at every time step.

Proposition 9. *Let $\mathbf{T}\alpha^k$ be a k-tape machine, with $k \in \mathbb{N}$. There exists a single-tape Turing machine $\mathbf{T}\alpha$ which simulates $\mathbf{T}\alpha^k$. More precisely, suppose that $\mathbf{T}\alpha^k$ begins with $x \in \Sigma^*$ on the first tape and with no symbols on the remaining $k-1$ tapes. If $\mathbf{T}^k_\alpha$ halts with $y \in \Sigma^*$ on the first tape, then, when $\mathbf{T}_\alpha$ is given x as input, it too halts with y as output.*

Outline of the proof. We know that the property holds when $\alpha = \omega_0$, as this reduces to the classic case of standard Turing machines. We seek to extend the same general proof idea to the case where the tape is of transfinite length.

We shall construct a machine $\mathbf{T}\alpha$ that simulates, on a single tape, the action of the k tapes of $\mathbf{T}\alpha^k$. To do so, $\mathbf{T}\alpha$ divides its infinite tape into k infinite sub-tapes by a staggered partition We represent this partition by a mapping ϕ from $\mathbb{N} \times \mathbb{O}$ to $\mathbb{O}$, which, given the pair (i, θ) representing the θ-th position of the i-th tape of $\mathbf{T}\alpha^k$, returns an ordinal θ' corresponding to a position on the tape of $\mathbf{T}\alpha$. The construction also encodes the positions of the read heads of $\mathbf{T}\alpha^k$. One simple method is to add a symbol μ_i at position $\phi(i, \theta)$ whenever the i-th read head of $\mathbf{T}\alpha^k$ is at cell θ. This symbol μ_k is chosen to be outside the alphabet Σ of $\mathbf{T}\alpha^k$. We also extend the alphabet Σ' of $\mathbf{T}\alpha$ by setting $\Sigma' = \Sigma \cup (\Sigma \times \mu_1, ..., \mu_k)$. Each position θ' on the tape of $\mathbf{T}_\alpha$ is thus either a symbol $\sigma \in \Sigma$ or a pair (σ, μ_i).

At each step, $\mathbf{T}\alpha$ will first simulate the behaviour of the first read head of the first tape of $\mathbf{T}\alpha^k$. To do this, it scans the positions $1, 1k+1, 2k+1, \ldots$ corresponding to that tape until it finds a pair (σ, μ_1). It then refers to the transition function δ to determine the next action. This function is constructed so that

Recall that $\mathbf{T}\alpha$ simulates $\mathbf{T}\alpha^k$ by encoding the contents of the k tapes of $\mathbf{T}\alpha^k$ on its single tape. One approach, as indicated in Figure **??**, is to extend the alphabet Σ' of $\mathbf{T}\alpha$ by defining $\Sigma' = \Sigma \cup (\Sigma \times 0, 1)^k$. Each cell on the tape of $\mathbf{T}\alpha$ then contains a $2k$-tuple of the form $(a_1, b_1, a_2, b_2, \ldots, a_k, b_k)$ with $a_1, ..., a_k \in \Sigma$ and $b_1, ..., b_k \in 0, 1$. This $2k$-tuple means that the first tape of $\mathbf{T}\alpha^k$ contains the symbol a_1, its second tape the symbol a_2, etc. Moreover, $b_i = 1$ indicates that the read head of the i-th tape is currently scanning the cell with symbol a_i, while $b_i = 0$ means that the read head is not scanning this cell. The details of this approach can be found in *Elements of the Theory of Computation*[9].

We know that the result holds for $\alpha = \omega_0$, since this is the case of ordinary Turing machines. We proceed by induction, and assume, for a certain non-inaccessible β, that if the property holds for β, then it holds for the smallest limit ordinal α with $\beta \subset \alpha$. In view of Proposition 8, we may assume, without loss of generality, that $\mathbf{T}\alpha^k$ only uses, for each of its k tapes, the segment between the symbol $\triangleright$ and $|\alpha\triangleright$. Suppose that the input x is placed on its first tape, and that after a certain number of operations, it halts with y written on that tape. If, at no point in the process, the symbol $|\alpha\triangleright$ is reached, then the machine behaves exactly as a machine of type $\mathbf{T}\beta^k$ (in particular, this means that strictly fewer than β characters of x have been read by the first read head). By the induction hypothesis, we can therefore construct a machine $\mathbf{T}\beta$ which simulates the behaviour of $\mathbf{T}\alpha^k$. It suffices to provide as input the β characters of x which have actually been accessed. From there, we can construct a machine $\mathbf{T}\alpha$ which is the α-extension of $\mathbf{T}\beta$. According to Proposition 5, this machine behaves exactly like $\mathbf{T}_\beta$ on any input x of length less than β. Thus, it will eventually halt with y as output.

Now suppose that, at some step τ in the computation, one of the read heads of $\mathbf{T}\alpha^k$ reaches the symbol $|\alpha\triangleright$ after γ operations. Whether the machine reaches $|\alpha\triangleright$ in the state ∞ or in a state $q \in K$, it receives a new state $p \in K$ (not necessarily distinct from q) and makes a move to the left (since, by Proposition 8, we have assumed that the machine does not use the cells beyond $|\alpha\triangleright$, a move to the right does not occur). The k-th read head thus now finds itself on the symbol $\triangleright$ in state p. By our induction hypothesis, our machine $\mathbf{T}\alpha$ can simulate the behaviour of $\mathbf{T}\alpha^k$ for all positions prior to $|\alpha\triangleright$. But at the moment when the k-th tape of $\mathbf{T}\alpha^k$ reaches $|\alpha\triangleright$, so too does $\mathbf{T}\alpha$. It therefore suffices to add at $|_\alpha\triangleright$ the rules which will tell it how to proceed in these circumstances (for instance, to reset to the initial symbol $\triangleright$ or adjust the state accordingly).

By iterating this strategy for each of the k tapes and for each possible transition, we ensure that every computational step of $\mathbf{T}\alpha^k$ is faithfully simulated by $\mathbf{T}\alpha$. Thus, a single-tape machine is as powerful as a k-tape machine, even in the transfinite case. □

[9] Lewis R. H. and Papadimitriou C. H., *ibidem*, p.206.

10.3 Comparison with existing littérature

10.3.1 Comparison with a Hierarchical Oracles Machine

There are two essential differences between the transfinite Turing machine we constructed and a Turing machine with an oracle. The first is that the oracle can be called in a single time unit. This allows one to decide, almost instantaneously, whether a machine M accepts a word x, whereas this operation could require ω_0 steps from a transfinite Turing machine. The second, and most important, is that the oracle can be called at any moment. It is enough to write the "question" one wishes to ask on its tape. In contrast, a transfinite Turing machine may loop ω_0-infinitely before reaching only the symbol $|_0\triangleright$. Furthermore, the machine does not allow one to answer any question. The only additional information it provides, compared to a classical Turing machine, is to determine whether the sequence of states that the transfinite Turing machine goes through converges to a specific state. This data includes, notably, the answer to the question of whether the machine halts, because if it were to halt, it would not reach the symbol $|_0\triangleright$. Thus, the machine learns, when it reaches $|_0\triangleright$, that it does not halt — but it learns this in ω_0 steps!

It is as if the machine $\mathbf{T}_\alpha$ could pass through different levels of meta-languages. At the first level, it behaves exactly like a classical machine and cannot resort to any form of oracle. At the second level, it gains the ability to determine the halting of certain types of operations performed at the first level, which will allow it, as we will show later, to transcend Rice's theorem. In concrete terms, this means that it has unlocked access to an oracle that can ask all sorts of questions about the semantics of the languages at that level. Little by little, a transfinite machine $\mathbf{T}_\alpha$ will thus unlock a series of successive oracles.

One might be tempted to compare a transfinite Turing machine to a classical Turing machine with a hierarchical oracle system. Let us denote such a machine by $\mathbf{T'}_i$, where i indicates the maximum level of the oracle that it is capable of unlocking. The oracle $\mathcal{O}_0$ takes as input a string x of arbitrary length (potentially infinite), as well as the encoding of a certain state q, which we will write as q^*. To distinguish the encoding of q from the string x, we can imagine that a reserved character, denoted \$, separates the concatenation of these two symbol sequences. $\mathcal{O}_0$ is then capable of telling whether the machine $\mathbf{T'}_i$, when started on input x, converges to q after a countably infinite number of operations. Formally, we say that $\mathcal{O}_0$ decides whether x belongs to the language L_q, formed by the set of words for which the state of $\mathbf{T}_1$ converges to q. In the special case where the encoding of q is the empty string, the oracle $\mathcal{O}_0$ tells whether the machine $\mathbf{T'}_i$ halts on input x (in finite time).

Higher-level oracles are then defined recursively. The oracle $\mathcal{O}_i$ takes as

input a string $x.\$.q^*$, where the "." denotes concatenation. This oracle is capable of telling whether the machine $\mathbf{T}'_i$, when started on input x, converges to q after a countably infinite number of operations, *and this without using oracles of strictly higher levels than i.* A positive answer then means that $\mathbf{T}'_i$ converges to q on x using only the oracles $\mathcal{O}_0, \mathcal{O}_1, ..., \mathcal{O}_{i-1}$. In the special case where q^* is the empty string, $\mathcal{O}_i$ answers the question of whether $\mathbf{T}'_i$ halts on x (in finite time), using only the lower-level oracles. Assuming that it is possible to simulate a machine $\mathbf{T}_{i-1}$ with a machine $\mathbf{T}'_{i-1}$, that is, a classical Turing machine equipped with the oracles $\mathcal{O}_0, \mathcal{O}_1, ..., \mathcal{O}_{i-1}$, the question answered by $\mathcal{O}_i$ is whether $\mathbf{T}_i$ loops ω_{i-1}-infinitely. We thus have a recursive strategy for proving the equivalence of transfinite machines and hierarchical oracle systems for any ordinal ω_i.

To understand how this approach would work in a real context — or why it might not — suppose that a machine $\mathbf{T}'_i$ is designed to mimic the behaviour of a given transfinite machine $\mathbf{T}_i$. Ideally, we would want that, on the same input x, both machines halt on the same outputs y. Moreover, we would need to relate the execution times of $\mathbf{T}_i$ and $\mathbf{T}'_i$ in a one-to-one fashion. The problem thus defined, the first thing $\mathbf{T}'_i$ would do when it is launched on input x is copy its input onto the oracle $\mathcal{O}_0$, using the empty string as the encoding for q^*. This oracle gives a positive answer if $\mathbf{T}'_1$ halts on x and a negative answer otherwise. If the answer is positive, machine $\mathbf{T}'_1$ then does exactly what $\mathbf{T}_{\omega_i}$ would have done on input x. Since $\mathbf{T}_{\omega_i}$ halts after a finite number of operations, Proposition 5 guarantees that its behaviour can be simulated by a classical Turing machine.

In the case where $\mathcal{O}_0$ answers negatively, however, the machine $\mathbf{T}'_1$ unlocks oracle $\mathcal{O}_1$ and starts it again on input x, without any encoding for q^*. By repeating the operation as many times as necessary, it either ends up on an oracle $\mathcal{O}_j$ that answers negatively, or exceeds the maximum oracle level i allowed. In the first case, this means that the machine loops ω_{j-1}-infinitely, but not ω_j-infinitely. In the second case, we conclude that $\mathbf{T}'_1$ loops ω_i-infinitely. This information is useful, as it already allows us to deduce the time complexity of the machine $\mathbf{T}_1$ simulated by $\mathbf{T}'_1$. But the second objective, let us recall, is to allow both machines, when they halt, to halt on the same output y. That is why $\mathbf{T}'_1$, as soon as it reaches an oracle $\mathcal{O}_j$ that answers negatively, will simulate the steps performed by $\mathbf{T}_1$ by adopting the following rule: each time $\mathbf{T}_1$ performs an operation, $\mathbf{T}'_1$ first copies the state of the tape in progress onto the oracle $\mathcal{O}_j$, using each possible state of $\mathbf{T}_1$ as the encoding for q^*. If any of these questions receives a positive answer, it means that $\mathbf{T}'_1$ converges to a certain state q. If not, it means that the machine diverges. $\mathbf{T}'_1$ can then deduce whether $\mathbf{T}_1$ will reach the symbol $\triangleright_i$ in state q or in state ∞. For convenience, we assume that $\mathbf{T}_1$ never uses locations beyond the first $\triangleright_i$, in the spirit of Proposition 8.

It is, however, difficult, with this single information about the conver-

gence of the machine's state, to exactly simulate the behaviour of $\mathbf{T}_1$. Before reaching the symbol $\triangleright_i$, in fact, a succession of transformations can be performed on the tape. Symbols may be added, others may be erased or replaced. We would like to truncate x to a certain finite size N, and decree that $\mathbf{T'}_1$ is capable of simulating the behaviour of $\mathbf{T}_1$ as N tends towards infinity. Nevertheless, this strategy does not, in general, guarantee the production of an identical output to the truncation of y.

For any fixed size N, it may indeed happen that symbols located beyond the N-th position of the tape cause changes to the symbols located upstream, without making the output indeterminate. An example of this phenomenon would be that of a machine $\mathbf{T}_1$ with the alphabet $\{a, b, c\}$ and having two states, q_1 and q_2. In state q_1, the machine progresses to the right, replacing all encountered characters with b, except for the symbol a, which causes the machine to enter state q_2. This state instructs the machine to :

1. Record its current position n;
2. Erase all characters between position n and the last symbol a encountered, which is itself erased (or, failing that, to proceed to the last symbol $\triangleright$ or $|_0\triangleright$);
3. Return to position n and transition to state q_1. When $\mathbf{T}_1$ reaches $|_0\triangleright$, it halts regardless of its state.

It is easy to see that such a machine will always halt on $|_0\triangleright$ in either state q_1 or q_2. Furthermore, the machine is stable, in the sense given by box ?? . However, it is not replicable by a hierarchical oracle system. Whatever integer N is chosen, it is always possible that a character a, or even an infinite number of a's, lies beyond N. Even if the oracle predicts that the machine will reach $|_0\triangleright$ in state q_1, this does not guarantee that no a will be encountered on the way (an even number of a's is enough to place the machine in state q_1).

In light of this discussion, reproducing the behaviour of $\mathbf{T}_1$ on an infinite input x using a hierarchical oracle Turing machine $\mathbf{T'}_1$ seems impossible in general, although it could be done on a truncation of x tending towards infinity. Yet, by virtue of Proposition 5, a machine $\mathbf{T}_1$ executed on a finite truncation behaves the same as a classical Turing machine. Thus, our hierarchical oracle Turing machine is unable to simulate a transfinite Turing machine. One might say, with caution regarding the meaning of words, that potential infinity cannot simulate the computational power of actual infinity, although a rigorous proof of this idea remains to be provided.

10.3.2 Comparison with a Continuous Turing Machine

Another thorny question consists in asking whether the transfinite Turing machine I have introduced, when all is said and done, might not resemble a

continuous Turing machine. And, should that be the case, what might my formalism contribute to such a model?

A continuous Turing machine, denoted **C**, is an analogue machine. Its tape is not divided into a countable infinity of cells, but instead appears as a truly continuous magnetic ribbon, the value of which may be read at any point in the space it occupies. Formally, it is described by a function $m(t, x)$ assigning a value to each point $x \in \mathbb{R}$ on the tape and to each instant t. There thus arises a certain resemblance between this continuous machine **C** and the formalism of transfinite Turing machines **T** that I have previously introduced. I shall begin by providing a more precise definition of a continuous Turing machine **C**. I emphasise a machine, not the machine, as there exist various competing definitions. I shall then illuminate several advantages of my own formalism.

Definition of a Continuous Turing Machine

For the purposes of our model, it is posited that the physical space in which our machine **C** is envisioned is a continuous space, in which each segment is infinitely divisible. It is further assumed that one may measure, with sufficient precision, the magnetic value at each point of this space. These assumptions may well come into conflict with quantum mechanics, or at least with a certain interpretation thereof, but such theoretical caveats do not, however, spoil the mathematical model we are constructing here, whose aim is to capture the notion of "continuous computation". I leave it to the reader to ponder whether such a machine might be "approximable" in our physical universe.

Our machine **C**, in addition to its continuous tape, is equipped with a read-head. This moves by following an acceleration $a(t)$ which may be controlled by the machine itself. Hence, it does not leap from one discrete cell to another, but traverses the tape in a continuous fashion. Its passage over the tape, and the reading of the magnetic field performed along its trajectory, may have the effect of accelerating or decelerating its motion. Under certain circumstances, this read-head may come to a complete halt—in which case we shall say that the machine has stopped. The magnetisation $m(t, x)$ of the tape after halting definitively determines the output s of the machine.

It is important to note that the machine's own states are represented too by a continuous function $e(t)$, whose value may change at every moment t, as a function of the magnetic value $m(t, x)$ presented by the tape at that instant, at the position occupied by the read-head. The function $a(t)$, governing acceleration, is thus dependent upon the current state $e(t)$ of the machine, which itself is a function of the read-head's position and of the tape's magnetisation at that point.

For the sake of completeness, let us note that our machine has the capacity to magnetise or demagnetise the value of the tape at $m(t, x + \delta)$, where

x marks the location of the read-head and δ is an arbitrarily small offset between where the head is reading the present value and where it actually writes a new value (so as to avoid the ambiguity that would result from the machine reading and writing the very same point simultaneously). It is reasonable to think that changing the value at one point on the tape also affects its adjacent positions. Such changes could be described in accordance with the established equations of electromagnetism[10]. Yet, in the context of a continuous Turing machine, we will nonetheless suppose that the machine is capable, with arbitrarily fine resolution, of changing the magnetisation at the point x without affecting neighbouring locations. This is a property one might term *infinite resolution*.

The continuous Turing machine herein described thus bears, broadly speaking, a resemblance to the old analogue video cassettes once manufactured (though the programme underlying those cassettes was, by contrast, very simple: it consisted merely in reading the tape from left to right at constant speed. Yet, we might well imagine a more sophisticated video recorder). On a purely mathematical level, it is crucial to ask in what way the continuous Turing machine surpasses the computational power of the classical Turing machine.

The Dead Ends of the Continuous Turing Machine

At first glance, we can readily see that our machine **C** possesses much more information, on its magnetic tape, than a classical Turing machine. This advantage, in terms of information storage, can only prove fruitful if the machine can indeed access such information. If, in fact, the machine does not, in the course of its operation, read every single point of its tape, or if such readings have no bearing on its subsequent behaviour, then those unused points might as well be altogether omitted. This is a subtlety of utmost importance for our discussion, as it leads to the following question: does the addition of storage genuinely increase the sorts of computations that a machine **C** can perform, compared to what a classical Turing machine could already achieve?

The fact remains that, in order to actually visit every single point on its tape, the continuous machine must be able to read and change the contents at each point instantaneously as the read-head traverses it. This property is what I term *infinite resolution*. For example, when the machine scans any given interval $[a, b]$ of the magnetic tape, it potentially modifies every point under its path. By Cantor's theorem[11], these points are uncountable. The

[10] For a simple and elegant description of this machine, as well as a glimpse at some of the issues it raises, see Kevin Brown's article (`https://www.mathpages.com/home/kmath135/kmath135.htm`).

[11] Cantor's theorem states that on every interval of the real line, the number of reals contained is uncountable. Hence, it strictly exceeds the number of natural numbers

machine thus performs an uncountable infinity of operations within each interval of time, however small.

Let us concede the possibility of infinite resolution—a concept that poses no fundamental dilemmas within the mathematical model—and let us then ask: does infinite resolution enhance the computational power of the machine? Certainly, our machine can read the value of a real function at every point, but a real function is nothing more than a relation between inputs and outputs. There is no need to compute the values of $f(x)$ at every point if one knows, for example, that $f(x) = x^2$. The pointwise value of the function is needed only when the relationship between input and output is unknown. Therefore, one might deem that the power of a continuous machine lies not so much in its ability to compute real functions, but in its capacity to manipulate vast sets of points between which the functional relationships are either unknown or highly complex.

It is possible to give a mathematical sense to this intuitive notion of "complexity" by considering it the minimum number of points necessary to specify the function. A linear function requires two points, and therefore is highly regular. A polynomial function of degree n requires $n+1$ points, and so forth. An arbitrary real function, provided it is *continuous*, may in fact be perfectly described by restricting both its domain and range to the rationals $\mathbb{Q}$. Now, the rationals are countably infinite. What I mean is that even though a continuous real function is defined over an uncountably infinite set, and one could thus evaluate it at an uncountable number of inputs, it suffices to describe its behaviour on a countably infinite set of points to characterise its whole action. Traversing the rationals $\mathbb{Q}$ is something well within the scope of a classical Turing machine (even if the latter cannot enumerate all rationals at once; we shall return to this distinction).

To circumvent this difficulty, perhaps one will admit that the continuous Turing machine can process discontinuous real functions. Such functions, indeed, cannot be wholly described by considering their behaviour merely over the rationals. The question that then arises is: how many discontinuities could such functions possess? If the number of discontinuities is finite, then this is little different to the situation in which the machine manipulates continuous functions; it suffices to manually encode the points of discontinuity, adding them, as it were, to the set of rationals over which the function is defined.

If, however, the discontinuities are infinite in number, then truly, we gain something more than for continuous functions. A famous example of functions discontinuous at infinitely many points is the Dirichlet function, which is constructed precisely to assign different values to points x from the domain, depending on whether x belongs to the rationals or to the irrationals. This function, denoted $1_{\mathbb{Q}}$ and also known as the *indicator function*,

---

$1, 2, 3, \ldots$.

is defined as follows:

$$\sigma(s,i) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

Barring such pathological cases, a continuous Turing machine typically does not wield more computational power than the classical Turing machine. In truth, the principal difference lies herein: our continuous Turing machine can, in principle, enumerate all the rational numbers of a line segment in finite time, whereas this would require a countably infinite time ω_0 for a transfinite Turing machine. This difference is what I have chosen to call *transcendence*. It consists in attaining the completion of an infinite enumerative process, and handling the outcome of this infinite process as a mental object in its own right. Is it not, however, precisely *transcendence* that characterises transfinite Turing machines?

A type $\mathbb{T}_1$ Transfinite Turing Machine, once expressed in normal form, is nothing other than a classical Turing machine endowed, as its sole additional property, with the capacity to access "the end" of its tape (transcendence). A machine of type $\mathbb{T}_2$, on the other hand, is capable, in addition to transcendence, of traversing an uncountable infinity of real numbers. It would thus appear that the $\mathbb{T}_2$ machine in my formalism corresponds to the continuous Turing machine. The $\mathbb{T}_1$ machine, conversely, would be a continuous machine capable only of traversing a countable infinity of reals. It could thus cover all rationals and some negligible proportion of reals, but, as it were, would exhibit "gaps" throughout its read-tape. However lofty its resolution, it does not yet attain *infinite resolution*.

Transfinite Machine vs. Continuous Machine

At the conclusion of this discussion, one must admit that the machine **C** and the transfinite machine of type $\mathbf{T}_2$ are strikingly similar. Each can enumerate real numbers—that is, each can "count" the uncountable infinite[12]. Both, moreover, share the property of *transcendence*.

It is true that the transfinite machine requires infinite time to "count infinity" on a ligne segment, whilst for the continuous machine this feat is dispatched in finite time. In this respect, the machine **C** appears more efficient. Yet this distinction concerning the duration of computation is, in the end, only a conceptual artefact. It is perfectly possible to define the

[12]More precisely, let us note that $\mathbf{T}_2$ is constructed to enumerate $\aleph_1$, whereas **C** is capable of enumerating the entirety of the reals. This set, by Cantor's diagonal argument, is equipotent to the power set of the integers, of cardinality $2^{\aleph_0}$. The question of whether $\aleph_1 = 2^{\aleph_0}$ is the famed continuum hypothesis, now known—thanks to Gödel and Cohen—to be undecidable within ZFC. Thus, today's standard axiomatisation of set theory provides no answer as to whether or not there exists a level of infinity strictly between the integers and the reals.

transfinite Turing machine so as to allow it to complete an infinite counting in finite time. The simplest means would be to posit that the time taken for it to perform an operation, such as advancing a step on the tape or reading/writing a value, decreases as the read-head advances along the tape. If we let δ_i denote the execution time taken by the machine when its read-head is on the ith position ($i = 0$ being the start cell), we might set for instance:

$$\delta_i = \frac{1}{2^i}$$

Thus, as the machine moves from left to right up to the end $\triangleright$ of the tape, the total elapsed time is

$$\sum_{i=0}^{\infty} \frac{1}{2^i} = 1$$

This convention is that adopted by the formalism of the Zeno Machine[13]. It grants the transfinite machine the power to count all natural numbers in finite time, as does the continuous Turing machine. Moreover, it is possible to generalise this approach to allow a machine of type $\mathbf{T}_2$ to count "all the real numbers" within finite time as well. To do so, one simply recalls that a machine $\mathbf{T}_2$ may be viewed as furnished with an infinite tape of size $\aleph_1$, itself decomposed into an infinity of sub-tapes of size $\aleph_0$. These tapes follow each other in an ordered sequence, so that we may suppose that the time needed for an operation decreases both within a tape, and as we shift from one tape to the next.

The reason for which I have not adopted this convention of shrinking computation times is that it is unnecessary. From a strictly mathematical point of view, it effects no change in the computational power of the machine. Anything computable by one in finite time may be done by the other in—at most—countably infinite time. In one instance, we might say the machine has taken, say, m computation steps, including three countings through infinity. In the other, one would count $\omega_0 \cdot 3 + 1$. There seems no reason to exclude this latter notation; indeed, it arguably possesses greater convenience.

In summary, the transfinite Turing machine may be seen as occupying an intermediate station between the classical Turing machine and the continuous Turing machine endowed with infinite resolution. The advantages of my description, when set against a continuous machine lacking infinite resolution, are multiple:

[13]The notion was first introduced and discussed in 1927 by Hermann Weyl. A more contemporary perspective may be found in *Zeno Machines and Hypercomputation* (Petrus H. Potgieter, 2006).

- **Minimalism:** A Continuous Turing machine $\mathbf{C}$ deprived of infinite resolution cannot access every real point on its tape. That is, though it possesses abundant information stored on the tape, it may only ever access a countable infinity of points. Thus, by an argument reminiscent of Occam's razor, it is preferable to opt for my transfinite Turing machine description. In this latter, the tape contains only a countable infinity of points, yet the machine ranges across its entire length. We have thus discarded points that will never be reached, while retaining the machine's truly essential property.

- **Generalisation:** The transfinite machine $\mathbf{T}_1$ which I have introduced may be generalised to higher-order machines $\mathbf{T}_2$, $\mathbf{T}_3$, and so forth. It so happens that the generalisation $\mathbf{T}_2$ corresponds exactly to the continuous machine $\mathbf{C}$ with infinite resolution. It effectively possesses a read-tape with cells indexed from zero up to ω_1, that is, an uncountable infinity of cells. This machine may run for countably infinite time, but also for uncountably infinite time—thus avoiding the earlier discussed problem concerning the number of lines of code. A machine $\mathbf{T}_2$ consists of a finite number of lines of code, but truly enjoys infinite resolution.

- **Transposition:** Finally, my formalism makes transparent the relations between a classical Turing machine, a continuous Turing machine of finite resolution, and one of infinite resolution. Property 5 in fact states that, insofar as a machine of level $\mathbf{T}_n$ runs for a time less than ω_{n-1}, it behaves as a machine of type $\mathbf{T}_{n-1}$. Conversely, if such a machine begins running for a duration ω_{n+1}, it acquires the same computational power as a machine of type $\mathbf{T}_{n+1}$. Accordingly, it is a straightforward matter to transpose a classical Turing machine to one of higher level, and vice versa. These machines will possess the same lines of code. The only differences that may arise concern minor qualifications: for example, an instruction such as "repeat this indefinitely" must specify: indefinitely for a countably infinite or an uncountably infinite duration? Aside from such notational transcriptions, which concern basically the precision of the language, the transposition requires no modification to the code or structure of the machine. To put it otherwise, the essential parameter distinguishing a classical Turing machine from a transfinite or continuous machine is the *computation time* (with the tape extending as far as the machine dares venture).

To conclude this comparison between the two machines, one would still need to determine, in detail, precisely which operations fall within the regime

of a transfinite Turing machine: can it compute second-order logic? Second-order arithmetic? Is it subject to incompletenesses or undecidabilities? Such a systematic comparison, however, extends beyond the scope of this endeavour, and would require a comprehensive review of the literature. For this reason, I shall content myself, in the following pages, with presenting some of the most significant results: the existence of a universal transfinite Turing machine, as well as the discovery of certain incomputability theorems reminiscent of Turing's original halting problem.

10.3.3 Comparison with the Infinite Time Turing Machine

In their seminal 1998 article, *Infinite Time Turing Machines*, Joel David Hamkins and Andy Lewis introduced a mathematical formalism akin to that which I set out in previous sections. Their approach is to generalise the classical Turing machine, endowing it with the capacity to execute a truly infinite number of operations. An Infinite Time Turing Machine (ITTM) with an alphabet $\Sigma = \{0, 1\}$ thus behaves precisely like its classical counterpart for all finite computations. Yet, as soon as the number of steps reaches a limit ordinal α—such as ω_0, or $\omega_0 2$, and so forth—the following behaviours are observed:

1. The read-write head returns to the initial $\triangleright$ position on the tape.

2. The machine enters a special state, q_∞, which arises uniquely at these limit stages.

3. For each tape cell c_i, one reviews the transfinite sequence of symbols written there before the machine has executed α operations. If this sequence converges towards 0, the cell c_i is assigned the symbol 0; in all other cases, the cell is marked with 1.

Conceptual Equivalence

There are two main differences between the ITTM as defined above and the Transfinite Turing Machine (TTM) I have introduced: (1) the TTM uses the symbol ? whereas the ITTM does not; (2) the ITTM's tape has cardinality $\aleph_0$, whereas the TTM may have tapes of higher cardinality.

The first distinction is of no essential importance. In truth, the absence of the symbol ? does not diminish the computational power of the machine. To see this, let us suppose we have both a Transfinite Turing Machine T_1 and an Infinite Time Turing Machine T' that possess identical sets of states, transition functions, alphabets, and halting criteria.

I will show that whenever the symbol ? appears on a cell c_i of T, it is possible to recover this information using T'. This is simply achieved by performing a Cantor-style duplication of the tape in T' so that each cell c_i

is paired with a duplicate c_i'. The instructions are adapted so that whenever a symbol is written to c_i, its opposite is written to c_i' immediately. Thus, the two cells always hold complementary values.

As a result, if c_i converges to 1, c_i' converges to 0, and vice versa. If c_i does not converge, neither does c_i', in which case both display 1. This demonstrates that the use of ? is inessential for discerning convergence in a cell.

The point (2), by contrast, is rather more significant. Hamkins and Lewis's ITTM restricts itself to a tape of length ω_0; my own transfinite machines allow for tapes whose length may be strictly greater. In other words, Hamkins and Lewis extend only the chronological memory of the Turing machine, leaving the spatial memory untouched. My proposal, on the other hand, contemplates the simultaneous—and proportionate—expansion of both.

This sets my model apart as more general. The ITTM is designed to carry out infinitely many operations on a countably infinite tape, while the Transfinite Machine can perform a countable infinity of steps on a tape of potentially larger size. Nevertheless, the two concepts coincide when the TTM's tape is restricted to cardinality $\aleph_0$.

In such a scenario, both machines exhibit perfectly analogous behaviours, save only that the ITTM's tape is of ordinal ω_0, while the TTM's tape is potentially extensible ad infinitum—after ω_0 cells, one encounters a new $|_{omega_0}\triangleright$ symbol and a further ω_0 cells, and so forth. This distinction vanishes if one adopts the normal form representation of the TTM: in that case, only the first segment of length ω_0 is ever utilised, as established in Proposition 8.

Thus, to summarise, the ITTM is equivalent to a Transfinite Turing Machine of cardinality $\aleph_0$, represented in its normal form.

Comparing the Results

Although the two concepts are ultimately equivalent in power, it is worth noting that Hamkins and Lewis, in their article, pursue different lines of enquiry. Their principal concern is to delineate, as sharply as possible, the computational potential of the ITTM.

One of the greatest merits of their contribution surely lies in their introduction of two now-fundamental notions: that of the *clockable* ordinal and the *writable* ordinal. An ordinal α is deemed clockable if there exists an ITTM which halts exactly at time α—that is, after a transfinite procession of steps numbering α itself. In contrast, an ordinal α is said to be writable if there is an ITTM which, starting from a blank tape, succeeds in inscribing a concrete representation of α (perhaps, for instance, its binary code), and then halts. At first glance, one might suppose these two concepts to be essentially coextensive; indeed, for all recursive ordinals, this is precisely

the case. Yet, upon reaching the Church-Kleene ordinal—the least non-recursive ordinal—the matter becomes increasingly delicate. The authors demonstrate that, although every writable ordinal is also inscriptible, there exist ordinals which, while writable, evade clockability and thus cannot be reached in the strict temporal sense.

Another striking contribution of their article is the exploration of which sets of real numbers can be decided or semi-decided by an ITTM. The tape of the machine, though broad, is capacious enough to accommodate a single real, but never an uncountable infinity of such objects. Nevertheless, it is possible to define a set of reals A by means of a criterion of the form $x \in A$ if and only if $\phi(x)$, where ϕ is an arithmetic predicate. Under this approach, decidability refers to the capacity to answer the question $\forall x, \phi(x)$, whereas semi-decidability concerns itself with the existence of some x such that $\neg\phi(x)$. The authors show that such questions fall within the purview of the ITTM. Indeed, they venture further, demonstrating that the machine is capable of deciding membership for even more intricately defined sets, such as those given by formulae of the type:

$$\forall x \exists y \phi(x, y)$$
$$\exists x \forall y \phi(x, y)$$

By contrast, the ITTM is unable to determine relations whose definitions involve an additional alternation of quantifiers, such as:

$$\forall x \exists y \forall z \phi(x, y, z)$$
$$\exists x \forall y \exists z \phi(x, y, z)$$

For comparison, a classical Turing machine is limited to determining only those relations on the natural numbers that involve a single quantifier—never two. Formally, one may express this by stating that ITTMs can decide the sets at level δ^1_2, while classical Turing machines are confined to the realm of δ^0_1.

Whilst these two results are of the highest interest, they do not correspond precisely to the questions I have attempted to address in my own approach. My attention is directed not toward the sheer computational power of the transfinite machine, but rather toward the precise relationships that may obtain between a Turing machine $\mathbf{T}_0$ (incapable of counting the infinite), a machine $\mathbf{T}_1$ (capable of ranging over the countable infinite), and a machine $\mathbf{T}_2$ (capable of enumerating the reals).

More precisely, I have asked whether, when one constructs an algorithm on a machine $\mathbf{T}n$ to resolve a given problem over the reals, it is possible to

transplant that algorithm to a machine $\mathbf{T}n-1$ in order to solve an analogous problem over the naturals. An affirmative response would suggest that the two machines manifest analogous structures; a negative answer would entail that, for a machine, the ability to count the uncountable infinite effects a fundamental transformation in its activity.

As will become apparent, it is this latter possibility that holds true. To establish this, I shall consider a generalised version of the 3-SAT problem. I will construct an algorithm that solves this problem "quickly" for a Transfinite Turing Machine—but which, when transposed to a classical Turing machine, yields only "slow" results. This suggests that accounting for the infinite is not a merely adjunct property, but one that profoundly alters the manner in which a machine reasons. I shall endeavour to show that this difference arises from the way in which the infinite licence a more global, holistic mode of reasoning—while without it, the machine is condemned to merely local operations, fragmented and isolated, each from the rest.

10.3.4 Universal Machines and the Generalized Halting Problem

Here, I will present some ideas for constructing a universal transfinite Turing machine. Let $\alpha > \omega_0$ be an ordinal, and let β be the largest limit ordinal such that $\beta \subset \alpha$. A Turing machine $\mathbf{T}_\alpha$ in normal form contains parts of its tape that are blank and remain blank throughout the process. These parts are equipotent to β and each has length β. Thus, it is possible to use these sub-tapes as full tapes. The first sub-tape would contain the information needed to encode the operation of another arbitrary Turing machine $\mathbf{T'}_\alpha$. Another sub-tape could be used to encode the state of $\mathbf{T'}_\alpha$ at the initial moment. A third would begin with an input x of size assumed to be less than α. A fourth, in turn, would start empty. At each unit of time, the machine thus obtained will read the content of the third sub-tape, considering it as the content read by the machine $\mathbf{T'}_\alpha$. Using the content of the first and second sub-tapes, our machine will then simulate the behaviour of $\mathbf{T'}_\alpha$ and report the input on the fourth sub-tape. This fourth sub-tape would thus contain, at the end of the process, the output given by $\mathbf{T'}_\alpha$ when executed on input x. Such a machine is denoted $\mathbf{U}_\alpha$. It will allow us to derive the following results.

Définition 20. *Let β be a limit ordinal. The generalized halting problem consists of deciding whether there exists a limit ordinal $\gamma \subseteq \beta$ such that $\mathbf{T}_\beta$ loops γ-infinitely.*

Theorem 8 (Generalized Halting Theorem)**.** *Let α and β be two limit ordinals, with $\alpha \subseteq \beta$. There exists no transfinite Turing machine $\mathbf{T}_\beta$ capable of deciding, in fewer than α steps, the generalized halting problem for a machine $\mathbf{T}_\alpha = (K, \Sigma, \delta, s, H, \alpha)$ and any input $x \in \Sigma^*$.*

Proof. Suppose there exists a machine $\mathbf{T}_\beta$ capable of deciding whether an arbitrary machine $\mathbf{T}_\alpha$ halts in less than α steps on some input x. Let $halts(\mathbf{T}_\alpha, x)$ denote the program thus defined, and $halts'(\mathbf{T}_\alpha, x)$ the reduction of this program on a machine of ordinality α. Proposition 5 ensures that $halts'$ and $halts$ exhibit the same behaviour for fewer than α instructions and on inputs smaller than β. By using an approach similar to the classical undecidability theorem, we could then construct the following program:

```
diagonal(x)
        if halts'(x, x)
                then loop alpha-infinitely
        otherwise
                halt
end
```

Note that the instruction "loop *beta*-infinitely" would consist of sending the reading head to the right of the tape until it reaches a symbol $|_\alpha\triangleright$.

We wonder whether *diagonal(diagonal)* loops α-infinitely? If not, then *halts'(diagonal, diagonal)* returns *true*, and the "loop beta-infinitely" instruction is launched, so that *diagonal* loops α-infinitely. But if it does, then *halts'(diagonal, diagonal)* returns $false$ and the *diagonal* program halts. However, it cannot halt in a time $\theta < \alpha$, as this would contradict the fact that *halts'(diagonal, diagonal)* finished with $false$. Therefore, the function *halts'(diagonal, diagonal)* must, although it returned $false$, have done so in a time $\theta \geq \alpha$.

We conclude that there is no function *halts'(x,x)* that halts in fewer than α steps. By corollary 1, this means there is also no program *halts* that halts in fewer than α steps. □

It is also possible to derive a generalized version of Rice's theorem.

Définition 21 (Non-trivial Property)**.** *A machine $\mathbf{T}_\alpha$ is said to possess a non-trivial property P if this property holds for every possible input of $\mathbf{T}_\alpha$, yet there exists at least one other Turing machine $\mathbf{T'}_\alpha$ for which there is at least one input that does not satisfy property P.*

Corollary 2 (Generalized Rice's Theorem)**.** *Let $\alpha \subseteq \beta$ be two limit ordinals. There exists no machine $\mathbf{T}_\beta$ capable of determining in a time $\theta < \alpha$ whether a machine $\mathbf{T}_\alpha$ satisfies a certain non-trivial property P on an arbitrary input x (a property is called "non-trivial" if there exists at least one input that satisfies it and one input that does not).*

Proof. General outline of the proof. This corollary can be demonstrated by proceeding in a manner analogous to the case of classical Turing machines. Suppose that there exists a programme A that is capable of determining, for

another programme, whether it possesses a certain non-trivial property. It is then shown that such a programme A would make it possible to determine, for any other programme, whether or not it halts, which contradicts the generalised halting problem. □

Définition 22. *Let β be a limit ordinal and Σ an alphabet. A language $L \subseteq \Sigma^*$ is β-enumerable (or β-decidable) if there exists a Turing machine $T_\beta = (K, \Sigma, \delta, s, H, \beta)$ that halts on all words $l \in L$.*

An important consequence of the generalized halting theorem is that there exist β-enumerable languages that are not β-decidable by Turing machines of ordinality $\alpha > \beta$.

Although a transfinite universal machine $\mathbf{U}_\alpha$ is subject to the same undecidability theorems as classical Turing machines *concerning the halting of programs of the same ordinality as itself*, one might wonder whether it is possible for $\mathbf{U}_\alpha$ to determine the halting of a machine of lower ordinality?

To clarify the question, suppose we have a universal machine $\mathbf{U}_\alpha$ and, without loss of generality, assume it has four sub-tapes at its disposal. $\mathbf{U}_\alpha$ can use the first three tapes to simulate the machine $\mathcal{T}_\beta$ on an input x (where $\beta \subset \alpha$ is a limit ordinal, and x is an input of size less than β). The fourth tape, however, would start empty, but each time $\mathbf{U}_\alpha$ simulates a new step of the operation of $\mathbf{T}_\beta$, it would be asked to add some arbitrary symbol $a \in \Sigma$ onto this tape. In this way, the size of the string written on the fourth tape would measure, at any moment, the execution time of $\mathbf{T}_\beta$ (in number of steps).

Let us specify that if the reading head of the fourth tape reaches a symbol of the type $|_\theta\triangleright$ (for some θ), it is instructed to move to the right and continue writing the symbol a as before. When the simulation of $\mathbf{T}_\beta$ finishes, one simply needs to look at the number of symbols on the fourth tape to determine how many total instructions were executed. This number may be finite, but it could also be an ordinal $\gamma \subset \alpha$. In the latter case, $\mathbf{U}_\alpha$ can unerringly conclude that the machine $\mathbf{T}_\beta$ it just simulated loops γ-infinitely on x. Moreover, we know that the simulation of $\mathbf{T}_\beta$ by $\mathbf{U}_\alpha$ is correct, by virtue of proposition 5.

The machine $\mathbf{U}_\alpha$ we have just described would thus be capable of solving the halting problem for any machine of lower ordinality β, but at most taking β iterations. Let us denote *loop0(M, x)* as the program that answers whether a machine M loops ω_0-infinitely on a finite input x, and apply it to the following program:

```
function1(x)
        write x to the tape
end
```

We expect that *loop0(function1, "a")* will return $false$ in a finite number of steps. So far, nothing is surprising. However, one may wonder what happens if we add, at the end of *function1*, the instruction *loop0(function1, x)*?

```
function1(x)
        write x to the tape
        loop0(function1, x)
end
```

In theory, the new *function1* should also halt in a finite number of steps, since the sequential execution of two finite programs ends in finite time. Yet, it is easy to see that *function1* will loop ω_0-infinitely. This is because the function *loop0(function1, x)* simulates *function1*, and *function1* in turn calls *loop0*, forcing it to simulate *loop0(loop0(function1,x),x)*, and so on. We conclude that a *loop0* function, constructed by blindly simulating the function it is supposed to measure the execution time of, may alter the time of execution itself. The measurement thus influences the result of the measurement.

Given this observation, we may ask whether there exists a way to construct a function *loop0(M,x)* that subtracts the effect of its presence from its result. In other words, we would like to know if a transfinite Turing machine can understand the semantics of any program of lower ordinality other than by simulating it? (or with a method faster than pure simulation). The theorem of generalized undecidability directly answers this question in the negative.

Other, more interesting results can still be derived.

Proposition 10. *Let β be a transfinite cardinal. The following properties hold:*

(i) There exists a machine $\mathbf{T}_{\beta+1}$ capable of determining, in a time equal to or less than $\beta + 1$, whether any machine $\mathbf{T}_\beta$ loops β-infinitely on any given input x.

(ii) There exists a machine $\mathbf{T}_{\beta+2}$ capable of determining, in a time equal to or less than $\beta + 2$, whether any machine $\mathbf{T}_\beta$ loops β-infinitely on all admissible inputs x.

(iii) There does not exist a machine $\mathbf{T}_{\beta+1}$ capable of determining, in a time equal to or less than $\beta + 1$, whether any machine $\mathbf{T}_\beta$ loops β-infinitely on all admissible inputs x.

Proof. (i) is obtained by constructing a universal machine $\mathbf{U}_{\beta+1}$ in normal form that simulates the behaviour of $\mathbf{T}_\beta$ and uses a blank sub-tape to count

the number of steps taken before halting (or reaching the symbol $|_\beta\triangleright$ on the sub-tape).

(*ii*) and (*iii*) require writing $\mathbf{T}_\beta$ in normal form and reasoning by contradiction. Notice that (*iii*) false would violate the generalized halting theorem (the proof is a bit lengthy). □

Proposition 11. *Let $\alpha < \beta$ be two ordinals, and let P be a property that is non-trivial for both machines of dimension α and machines of dimension β. The possession of property P by a machine $\mathbf{T}_\alpha$ is not equivalent to the possession of P by its β-extension $\mathbf{T}_\beta$.*

Proof. Suppose, for the sake of contradiction, that the following hold:

1. If an arbitrary machine $\mathbf{T}_\alpha$ possesses the non-trivial property P, then all its β-extensions $\mathbf{T}_\beta$ also possess P.

2. If an arbitrary machine $\mathbf{T}_\beta$ possesses the non-trivial property P, then its α-restriction $\mathbf{T}_\alpha$ also possesses P.

Under these assumptions, let us show that one could construct a machine $\mathbf{U}_\beta$ capable of determining, for any machine $\mathbf{T}_\beta$, whether it possesses property P. Since P is non-trivial, this would constitute a violation of Rice's theorem.

To construct such a machine $\mathbf{U}_\beta$, it would suffice to design an algorithm that verifies whether the α-reduction $\mathbf{T}_\alpha$ of $\mathbf{T}_\beta$ satisfies property P. This algorithm would simply run $\mathbf{T}_\alpha$ on all possible inputs and check whether there exists an input for which P fails to hold (which is possible, since $\alpha < \beta$).

If property P holds for all inputs, we may then deduce that $\mathbf{T}_\beta$ also possesses property P (by assumption 1). Conversely, if there exists an input for which $\mathbf{T'}_\alpha$ does not satisfy P, then $\mathbf{T'}_\beta$ too fails to possess property P (by the contrapositive of assumption 2). Thus, the two assumptions cannot be simultaneously maintained.

□

Example 2. *Let $\mathbf{T}_{\omega_1}$ be a machine that solves the 3-$SAT_{\aleph_0}$ problem for any valid input x. An interesting property P would be to ask whether $\mathbf{T}_{\omega_1}$ halts after a number of steps polynomially dependent on the size of the input x. In the case where this input is of length ω_0, this amounts to asking whether $\mathcal{P} = \mathcal{NP}$. In the case where the input has length ω_1, this amounts to asking whether 3-$SAT_{\aleph_0}$ is solvable in a countably infinite number of steps. We will show in the next section that this is a non-trivial property. In light of the above theorem, this means that the solvability of 3-$SAT_{\aleph_0}$ does not generally solve $\mathcal{P} = \mathcal{NP}$. However, one can circumvent this difficulty by imposing a*

particular relation between $\mathbf{T}_{\omega_1}$ *and its reduction* $\mathbf{T}_{\omega_0}$*. This is the spirit of the following proposition.*

There are several ways in which one might circumvent the hypotheses of Proposition 11. To begin with, it should be noted that the proposition holds only for those properties P which remain non-trivial irrespective of the dimension of the machine under consideration. This entails that there must exist at least one machine of dimension α which satisfies P, and at least one machine of dimension β which does not. Moreover, it is also required that among the machines of dimension β, there exist both a machine which possesses property P and one which does not.

In practice, however, the property asserting that a problem is solvable in polynomial time may well be non-trivial for transfinite machines, yet become a trivial property in the case of machines of finite size. Such a situation would arise, for example, if one had $P \neq NP$ in the finite case, but $P = NP$ in the infinite case.

Another way to circumvent Proposition 11 consists in supposing that one may infer certain facts about infinite machines by examining the behaviour of finite ones—though not, conversely, from infinite to finite. This amounts to accepting hypothesis 1 as presented in the proof of Proposition 11, whilst rejecting hypothesis 2. Symmetrically, one might accept hypothesis 2 but reject hypothesis 1[14].

A final approach is simply to recall that, although verifying a non-trivial property P on transfinite machines does not, in general, afford any significant conclusion regarding the verification of P on finite machines, exceptions can conceivably exist. It is not impossible, for instance, that a specific algorithm, or an entire class of algorithms, retains certain properties when transposed from the transfinite context to the finite.

Thus, it nonetheless remains possible, *a priori*, to shed some light upon the question $\mathcal{P} = \mathcal{NP}$ by way of transfinite Turing machines, although this would require determining a function establishing a correspondence between a subspace of the set of classical Turing machines and a subspace of their transfinite extensions (this is, in essence, the rationale behind the properties P_α and P_β). In fact, one might reasonably expect to find an injection from one to the other, or vice versa, rather than a bijective correspondence.

[14] I am inclined to believe, however, that this latter approach is not tenable. My view is that it is possible to introduce conventions—drawing, for instance, on the normal form of a Transfinite Turing Machine—such that any Transfinite Turing Machine of cardinality α admits one, and only one, β-extension. (As justification, observe that the Time Turing Machine formalism embodies the notion of a single algorithm that may be executed in finite time, in time ω_0, in time ω_1, and so forth; my intention is to establish an analogous arrangement, albeit adapted to the fact that my own machines admit tapes of even greater length.) Once this is granted, it becomes clear that hypothesis 1 entails hypothesis 2, and vice versa, thereby strengthening Proposition 11.

What to take away from this chapter

It is important to distinguish between two concepts:

1. A *function* is a black box that associates an answer to a question, or an output to an input.
2. The model is a description of the black box, explaining its internal mechanism.

The abacus, for example, is a *model* of arithmetic. By manipulating the abacus, one systematically finds the answers to the *functions* "addition" and "subtraction." However, although the abacus is a model of arithmetic, and arithmetic presents functions that the human mind is capable of, the abacus is not a model of the human mind.

In the same way, the computer is a *model* of all computable functions, but even though these functions are within the reach of the human mind, the computer is not a model of the human mind. The question that arises is to understand, therefore, what non-computable functions the human being possesses in addition to the computer?

In this chapter, I have considered three non-computable functions:

1. The completion of infinite series.
2. The resolution of undecidable problems for classical Turing machines.
3. The rapid resolution of NP-complete problems.

I have shown that functions (1) and (2) cannot be executed by a classical Turing machine, but by a transfinite Turing machine. In the next chapter, the aim is to show that the same is true for function (3).

Chapter 11

Generalised NP-complete problems

In the preceding chapter, I argued that intuition could be envisioned as a non-computable function, possessing a clear definition but not executable by a classical Turing machine. I explored three ways to construct this function: (1) seeing intuition as the ability to count the infinite, or more rigorously, as the skill to complete an infinite series; (2) viewing intuition as the capability to solve undecidable problems, for which no generic algorithm exists; (3) finally, as the ability to find, in a reasonable amount of time, a solution to an NP-complete problem of restricted size.

At first glance, these three strategies correspond to somewhat distinct visions of human intuition. Strategy (1) views intuition as linked to the presence, within us, of the idea of the infinite; because we can manipulate concepts of infinity and provide them with meaning, we do not function like machines. Strategy (2) sees intuition as the skill to resolve problems that resist logical deductive reasoning, implying a tendency within the human mind to reason otherwise than according to logical laws. Strategy (3) perceives intuition as the exercise of an inner voice, an unconscious guide, allowing us to have "luck" with certain notoriously difficult tasks. The presence of these sudden insights does not contradict the representation of the mind functioning like a Turing machine. However, it supports the idea of an additional, complementary process concerning our way of choosing an effective starting point for our reasoning.

I preferred strategy (1) and constructed a transfinite Turing machine. This preference stems from the fact that the ability to count the infinite closely resembles the definition of transcendent intuition derived in the tricameral theory of mind (see section 8.4). To recap, transcendent intuition is indeed the ability to comprehend that a concept such as a "tree" conceals an infinity of possible trees, resembling and partially differing from each other. The concept of a "tree" is conceived as the limit of a comparison process

where each language term refers to a multitude of other terms. Thus, in my view, there is a link between the intuition of infinity and the construction of universal abstract concepts.

Building a machine satisfying property (1) was no easy task. Through the sweat of my mathematical brow, I extended the computing power of the classical Turing machine, enabling it to perform an infinite number of operations. Concretely, I allowed it to reach the "end" of its allocated tape to read the "last" term of an infinite series. Of course, I defined the "end" of the tape and the "last" term inspired by Cantor's infinity theory; otherwise, these expressions would be meaningless. This led me to construct not just one transfinite Turing machine but an infinity of machines $\mathbf{T}_{\omega_1}, \mathbf{T}_{\omega_2}, ...$ The machine $\mathbf{T}_{\omega_1}$, also known as the first-order machine, can enumerate the infinity of integers; the machine $\mathbf{T}_{\omega_2}$ can count the infinity of real numbers, the machine $\mathbf{T}_{\omega_2}$ can count all infinite possible sets of real numbers, etc.

Like an architect who, after laying the foundations of a tower on paper, goes to the site to study the gradient, I sought to explore some of the concrete properties of this new mathematical formalism. I particularly endeavoured to determine whether undecidable problems existed for a transfinite Turing machine. To this end, I began with the halting problem conceived by Alan Turing, defined as follows: "Is there an algorithm to determine whether any Turing machine T halts in finite time on a given input s?" I demonstrated that it is possible to construct a transfinite Turing machine to solve the halting problem for a classical Turing machine. However, there is no algorithm for a transfinite Turing machine to determine if another transfinite machine of equivalent or higher order halts on an input s. In other words, the halting problem is extendable to transfinite Turing machines.

From these considerations, an infinite hierarchy of undecidable problems emerges. Some problems are undecidable for classical Turing machines, others for transfinite Turing machines of order 1, then for machines of order 2, and so on. Consequently, my formalism of transfinite machines allows both counting the infinite and solving certain classes of undecidable problems. In other words, my tool, initially designed for strategy (1), also serves well for strategy (2). Ultimately, this implies that the concept of intuition as the presence of an idea of infinity within us coincides with the notion that we might resolve, intuitively, tasks insoluble by a classical Turing machine.

In this chapter, I wish to explore two unresolved questions. The first is whether transfinite Turing machines can also shed light on strategy (3). In other words, I ask whether a transfinite machine is capable of finding a swift solution to an NP-complete problem, known for being laboriously complex to resolve. If this is the case, it implies that the three visions of human intuition we identified closely coincide. In technical terms, I will assert that these are three coextensive functions. Although they can be depicted separately, the model of transfinite machines, intended to express one of them, supports the other two. They are three aspects of a broader

general function.

Subsequent to this, and here lies the second question I wish to address in this chapter, it might be that holistic intuition, which I discussed in the tricameral theory of mind, is also coextensive with the other definitions of intuition previously given. To recap, this definition proposed that the human mind can apprehend information holistically, decomposing it according to determined perceptual structures, rather than processing it as a succession of binary units awaiting linkage.

To phrase this thesis succinctly: intuition, this multi-headed hydra, might be but a singular multifaceted monster; that a strong conceptual connexion exists between the conception of infinity, the holistic understanding of perception seen in optical illusions, and the resolution of certain undecidable or complex problems—such as NP-complete problems.

Definitions	Functions
Transcendent Intuition	Completing infinite series Solving undecidable problems
Holistic Intuition	Solving NP-complete problems Subject to perceptual illusions

In the following pages, I will first reiterate the definition of NP-complete problems and demonstrate why solving these generally requires exponential time. I will then show that transfinite Turing machines are no more efficient at these problems than a classical Turing machine. The reason being that a transfinite Turing machine behaves exactly like a classical Turing machine when asked to perform a finite number of reasoning steps.

However, I will construct a generalisation of NP-complete problems such that, for a transfinite machine, finding a solution to these problems is no more difficult than verifying if a given answer is a solution. This generalised NP-complete problem in question is nothing but an NP-complete problem involving continuous functions over the real line. Thus, it is a problem defined over $\mathbb{R}$ and not $\mathbb{N}$. This problem includes an infinite number of variables but corresponds to geometric problems that can be intuitively represented. If this notion of a generalised NP-complete problem still seems vague to the reader, I intend to clarify it in the following sections.

11.1 Classical NP-Complete Problems

As discussed in section 6.6, NP-complete problems form a single class with an equivalent structural core. This means that any resolution algorithm that can solve a problem P_1 from this class may, with a simple adaptation, be used to solve another problem P_2 from the same class. By "simple adaptation", we mean that the transformation of the algorithm can be carried out in

polynomial time—in other words, the computational time required to adapt the algorithm is negligible in comparison to the time needed for its actual execution.

This essential property makes it possible to select one problem from the NP-complete class to represent them all. The usual choice falls on the 3-SAT problem, which involves solving a logical formula. For convenience, this formula can be written in conjunctive normal form. This means expressing the formula as a set of clauses of the form "A or B". One might, for instance, consider clauses such as "A or not B", "not A or B", and "not B or not A", giving rise to the following formula F:

$$(A \vee \neg B) \wedge (\neg A \vee B) \wedge (\neg B \vee \neg A)$$

It is easy to verify that this formula admits a single solution, namely both A and B being false. For if A were true, then the third clause tells us that B must be false. But if B is false, the second clause implies that A must be false. This leads to a contradiction, from which we conclude that A must be false. Symmetrically, suppose that B is true. The third clause teaches us that A is false. But if A is false, then the first clause implies that B must also be false—contradicting the assumption that B was true. It follows that B must be false as well. In other words, one must assign the values $A = B = 0$.

This reasoning is confirmed in the truth table below. To recall, $\wedge$ and $\vee$ are the logical quantifiers "And" and "Or" respectively, while $\neg$ represents negation:

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

This problem is NP-complete, meaning that verifying whether an assignment of the variables A and B is a solution can be done quickly, while finding the correct assignment is likely to take significantly longer. More precisely, we may say that the number of steps required to verify that $A = B = 0$ is a solution involves checking each clause one by one. There is thus a linear relationship between the number of clauses and the time needed for verification. On the other hand, finding the solution $A = B = 0$, if done naively, would require going through all possible assignments and testing them one by one. In the case of two variables A and B, there are only four possibilities:

- $A = B = 1$
- $A = 1$ et $B = 0$

- $A = 0$ et $B = 1$
- $A = B = 1$

If the problem involved three variables A, B, and C, the reasoning would take eight steps. With four variables, it would require sixteen, and so forth. In the presence of n variables $v_1, v_2, ..., v_n$, one must perform 2^n operations. However, verifying that a proposed candidate satisfies the clauses never requires more than n operations, since one merely needs to go through each clause sequentially. It is easy to observe that 2^n is an exponential expression, while n is a polynomial one. From this observation arises the definition of an NP-complete problem:

Définition 23. *An NP-complete problem is one for which verifying a solution can be performed in polynomial time, while discovering a solution requires, at minimum, polynomial time, and at most, exponential time.*

NP-complete problems are omnipresent: when you try to solve Minesweeper, when you pack items into your hiking backpack, when you distribute tasks before a family dinner, when you pair up students while considering their preferences, when you seek to optimise a route through several shops—or any time you attempt to solve a problem that can be written in the form of a logical equation—you are facing an NP-complete problem.

No one knows whether a general resolution method exists that would allow one to find the solution to an NP-complete problem in polynomial time. All known methods exhibit weaknesses; no matter how robust the approach, there are always scenarios in which it requires approximately 2^n steps. In technical terms, we say that it is still unknown whether NP-complete problems belong to the class P—the class of problems solvable in polynomial time. We write $P = NP$ if that is the case, and $P \neq NP$ otherwise.

Experience suggests that, when confronted with such problems, the human mind does not always proceed via strict logical reasoning, but rather by taking a leap of faith. In the case of Minesweeper, for example, one quickly realises that the decision to click on a specific square demands an analysis of the neighbouring squares and, by gradual expansion, the entire game board. There is a global relationship between the choice to click the upper-left square and the information gleaned from squares located far away. Given the limits of our information-processing capabilities, we resort to formulating reasonable hypotheses, recognising recurring patterns, and so forth. The world record is 1 second for beginner level (8×8 grid), 8 seconds for intermediate, and 29 seconds for expert level (31×16 grid). Needless to say, such speed cannot be achieved solely through logical analysis (Kahneman's System 2), which is slow, cold, and sequential.

The question we must therefore ask is whether intuition might be characterised as a certain ability to swiftly identify a good candidate solution to this kind of problem. Given that the formal definition I have proposed for intuition is the transfinite Turing machine, the question becomes whether such a machine would enjoy any advantage, in solving an NP-complete problem, over the classical Turing machine. If it could be shown that the recourse to the infinite allowed for a better grasp of the global relations between the problem's variables, and thus for an accelerated resolution process, we would then hold an argument demonstrating a genuine connexion between the two definitions of intuition explored so far: the intuition of the whole, and the intuition of the infinite.

Alas, it takes little effort to realise this is not the case. A transfinite Turing machine differs from its classical counterpart only when granted an infinite time for reflection. So long as the number of steps executed by the transfinite machine remains finite, Proposition 5 teaches us that the transfinite machine cannot be distinguished from a classical Turing machine. This implies that both have the same algorithmic efficiency.

Does this result irrevocably compromise the overall approach? Not necessarily. It is true that the classical and transfinite Turing machines have the same efficiency when facing NP-complete problems of finite size. But since the transfinite Turing machine is an infinite extension of the classical one, the more pertinent question becomes: how does it behave when confronted with infinite extensions of NP-complete problems? Until now, we have restricted ourselves to problems involving a finite number of variables $v_1, v_2, ..., v_n$. But if this number were to tend toward infinity, how effective would our algorithms remain?

For a classical Turing machine, it is logically expected that the time required to verify a solution tends toward $\lim_{n\to\infty} n$, while the time required to discover a solution, using currently known algorithms, would tend toward $\lim_{n\to\infty} 2^n$, so that the ratio between the two yields:

$$\lim_{n\to\infty} \frac{2^n}{n} = 0 \tag{11.1}$$

This result, if we pause to reflect on it, appears quite reasonable. As the number of variables tends toward infinity, it requires infinite time to verify that a given solution satisfies the problem. Conversely, discovering a solution—given the algorithms currently available—would amount to checking every possible one. One must therefore verify, an infinite number of times, solutions whose verification itself requires infinite time. In George Cantor's theory, we would say that the verification takes a countably infinite time, denoted $\aleph_0$, while the discovery of a solution demands an uncountably infinite time, expressed as $2^{\aleph_0}$. And one easily sees that $2^{\aleph_0}$ is a greater infinity than $\aleph_0$.

In the sections that follow, I shall dismantle a counter-intuitive theorem:

for transfinite Turing machines, it is possible to construct a resolution algorithm such that, if the NP-complete problem indeed involves an infinite number of variables, then the ratio given by equation 11.1 equals 1. In other words, it is not significantly more difficult for a transfinite Turing machine to solve an NP-complete problem than to verify that a proposed answer constitutes a solution. Put differently, the results I will present suggest that the class NP^T of transfinite NP-complete problems is equivalent to the class P^T of transfinite polynomial problems. One may therefore write: $NP^T = P^T$.

11.2 Problem 3-SAT$_{\aleph_0}$

To examine the algorithmic efficiency of transfinite Turing machines, one must first generalise NP-complete problems to an infinite number of variables. As previously noted, the most well-known NP-complete problem is the 3-SAT problem, which consists in solving a logical equation. I will extend this 3-SAT problem to a countably infinite number of variables.

Let us consider a countably infinite set $\eta \subset \omega_1$ of **propositional variables D** (with $\eta \subset \omega_1$). In what follows, we shall index them—without loss of generality—as $v_1, v_2, \ldots$. This does not imply that $\eta = \omega_0$, but simply that the variables can be placed in bijection with the natural numbers. Let us also consider a countably infinite set $\mu \subset \omega_1$ of **clauses** of the form $l_i \vee l_j \vee l_k$ (with i, j, k distinct natural numbers). Each **literal $\mathbf{l_n}$** appearing in this expression may either denote the propositional variable v_n (in which case we speak of a positive literal), or its negation $\neg v_n$ (a negative literal). The **3-SAT$_{\aleph_0}$** problem consists in finding an assignment of propositional variables that satisfies the entire set of clauses.

We shall demonstrate that this problem is solvable through a method requiring, at most, a countably infinite number of steps. To this end, a geometric representation of 3-SAT$_{\aleph_0}$ will first be proposed, wherein the domain of admissible solutions will be assimilated to the real interval $[0, 1]$. Next, the clauses will be expressed by means of **harmonics** Ψ that vanish at certain key values of $[0, 1]$. Finally, a resolution algorithm will be introduced. For computational efficiency, this algorithm will be formulated independently of the geometric representation given above; however, the latter will serve to prove the convergence of the algorithm with maximal clarity and elegance. In conclusion, we shall show that the algorithm—written in pseudocode—can be implemented by a **transfinite Turing machine $\mathbf{T}_{\omega_1}$**, as defined in Section 10.2.

11.3 Geometric Representation of the Problem

It is possible to endow 3-SAT$_{\aleph_0}$ with a geometric representation. Instead of assigning the values *true* or *false* to the propositional variables $v_1, v_2, \ldots \in$

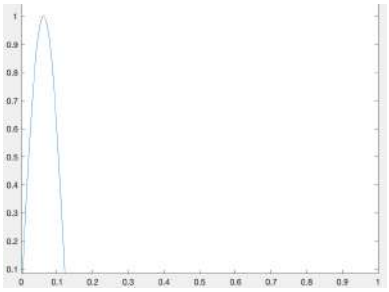
V, we may define a valuation ν from D into $\{0,1\}$. A variable $v_i \in D$ shall be said to be true if and only if $\nu(v_i) = 1$, and false when $\nu(v_i) = 0$. The space of admissible solutions for 3-SAT$_{\aleph_0}$ is thus a Cantor space: $\{0,1\}^{\mathbb{N}^*}$. Each element of this space can be identified with the binary expansion of a real number in the interval $[0,1]$. This is done by defining a function ϕ which associates the sequence $(v_i)_{i\in\mathbb{N}^*}$ with the number $\sum_{i=1}^{\infty} \nu(v_i)/2^i$. The sequence "101010…", for instance, would be represented by $2/3$, and "010101…" by $1/3$.

Within this framework, finding a solution to 3-SAT$_{\aleph_0}$ becomes equivalent to finding a real number on the interval $[0,1]$ that satisfies the clauses $l_i \vee l_j \vee l_k$ (with i,j,k distinct natural numbers). Now, as Cantor taught us, the cardinality of any segment of the real line is $2^{\aleph_0}$. Consequently, a machine capable of executing $\aleph_0$ operations could not naively scan through the admissible solution space, nor even through subintervals of the segment (excepting singletons of the form $[a,a]$, which contain only a single point). Unless there exists an uncountable infinity of solutions to the problem, the probability of stumbling upon a correct answer by chance is strictly equal to zero[1].

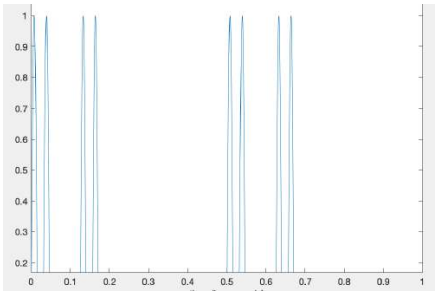
A meticulous reader might point out that the function ϕ defined above is not an injection from $\{0,1\}^{\mathbb{N}^*}$ into $[0,1]$. Indeed, sequences ending in an infinite string of 1s or 0s have the same image. For instance, "0111…" and "1000…" both map to $1/2$, yet represent different sequences—and, crucially, different *limits of sequences*. Suppose, for example, that we construct a countably infinite sequence of sequences $v_0, v_1, \ldots$ such that $\phi(v_i) < 1/2$ (for all $i \in \mathbb{N}$) and $\lim_{n\to\infty} \phi(v_n) = 1/2$. However we choose them, the sequences in this family will tend towards the sequence "0111…", which begins with a 0 followed by an ever-increasing number of consecutive 1s. Conversely, if we construct a sequence $v_0, v_1, \ldots$ such that $\phi(v_i) > 1/2$ and $\lim_{n\to\infty} \phi(v_n) = 1/2$, then this sequence will tend towards "1000…", beginning with a 1 followed by an ever-increasing number of consecutive 0s. Put in more intuitive terms, one may say that a point in the image set of ϕ may be approached from the left and from the right in qualitatively different ways. Yet this singularity arises only for certain specific rationals. And since the rationals are negligible within the set of real numbers, one may assert that ϕ is injective *almost everywhere* on the interval $[0,1]$.

Having made this methodological detour, the question we must now ask is whether the clauses of 3-SAT$_{\aleph_0}$ are rich enough in information to enable us to find a solution in, at most, a countably infinite number of steps. Rather than seeking the assignments $\nu(v_i)$ that satisfy the set of constraints, we might ask which assignments *falsify* them. Indeed, the dual problem is sometimes easier to visualise than the original. By applying De Morgan's

[1] Let us recall that in the context of sets of infinite cardinality, events of probability zero are not necessarily mathematically impossible.



(a) $\neg(\neg v_1 \wedge \neg v_2 \wedge \neg v_3)$



(b) $\neg(\neg v_1 \wedge \neg v_3 \wedge \neg v_5)$

Figure 11.1: Graphical representation of the intervals excluded by two different clauses. Each clause is associated with a periodic function $\psi : [0,1] \to \mathbb{R}$ such that if $x \in [0,1]$ is excluded by the clause, then either $\psi(x) > 0$, or $\psi(x) = 0$ and $\psi'(x) > 0$.

law, each clause $l_i \vee l_j \vee l_k$ is equivalent to $\neg(\neg l_i \wedge \neg l_j \wedge \neg l_k)$. A clause thus rewritten excludes certain subintervals of $[0,1]$ from the admissible solution space. The clause $\neg(\neg v_1 \wedge \neg v_2 \wedge \neg v_3)$, for example, essentially states that the solution cannot lie within the interval $[0, 1/8)$. In fact, the point $1/8$ is excluded if $(v_i)_{i\in\mathbb{N}} = 000111\ldots$, and included if $(v_i)_{i\in\mathbb{N}} = 001000\ldots$ This observation leads us to the following definition.

Définition 24. *Let two real numbers $a, b \in [0,1]$, a sequence $v \in \{0,1\}^{\mathbb{N}^*}$, and the function $\phi : \{0,1\}^{\mathbb{N}^*} \to \mathbb{R} : (v_i)_{i\in\mathbb{N}^*} \to \sum_{i=1}^{\infty} \nu(v_i)/2^i$. An exclusion interval, denoted $|a,b|$, denotes a subset of Dom(ϕ) such that $v \in |a,b|$ if and only if one of the following three conditions holds:*

1. *$\phi(v) \in (a,b)$.*
2. *$\phi(v) = a$ and v does not end in an infinite string of 1s. In this case, a is called a left limit point of the interval.*
3. *$\phi(v) = b$ and v does not end in an infinite string of 0s. In this case, b is called a right limit point of the interval.*

Example 3. *The clause $\neg(\neg v_1 \wedge v_2 \wedge \neg v_3)$ has the exclusion interval $|2/8, 3/8|$. We say that $2/8$ and $3/8$ are left and right limit points (respectively). The measure of this interval is $1/8$ (see figure 11.1a).*

Example 4. *The clause $\neg(\neg v_1 \wedge \neg v_3 \wedge \neg v_5)$ has the exclusion interval $|0, 1/32| \cup |2/32, 3/32| \cup |8/32, 9/32| \cup |9/32, 10/32| \cup [1/2, 17/32| \cup |18/32, 19/32| \cup |24/32, 25/32| \cup |26/32, 27/32|$. This interval is the union of eight disjoint intervals of measure $1/32$ (see figure 11.1b).*

Proposition 12. *For all real numbers a, b, c, and d, the following three expressions are equivalent:*

1. $|a,b| \cup |b,c| = |a,c|$;
2. $|0,a| = [0,a|$;
3. $|a,1| = |a,1]$.

Proof. A sequence $v \in \{0,1\}^{\mathbb{N}^*}$ satisfies $\phi(v) = 0$ if and only if it consists solely of 0s. It satisfies $\phi(v) = 1$ if and only if it consists solely of 1s. This proves points 1 and 2.

For point 1, suppose that $v \in \{0,1\}^{\mathbb{N}^*}$ satisfies $\phi(v) = b$. In this case, either v does not end with an infinite number of 0s, and thus $\phi(v) \in |a,b|$, or v does not end with an infinite number of 1s, and thus $\phi(v) \in |b,c|$ (this is an inclusive disjunction). In both cases, $\phi(v) \in |a,c|$. The cases where $\phi(v) \neq b$ are trivial. □

The above proposition guarantees that if one clause excludes the interval $|0,1/8|$ and another excludes $|1/8,1/4|$, we can conclude that the solution does not lie within the interval $[0,1/4|$. Thus, it seems possible to iteratively restrict the space of admissible solutions. However, we must determine which intervals are excluded by any given clause $\neg(\neg v_i \wedge \neg v_j \wedge \neg v_k)$. Ideally, we would like to describe these intervals using an analytical expression that is easy to manipulate.

Définition 25. *Let a clause $\neg(\neg l_i \wedge \neg l_j \wedge \neg l_k)$ be given in the 3-SAT$_{\omega_0}$ problem, with $i < j < k$. Define $\rho_i, \rho_j, \rho_k \in \{0,1\}$ such that, for each $m \in \{i,j,k\}$, $\rho_m = 0$ if the literal l_m is positive, and 1 otherwise. We represent this clause by the sextuplet $C = (i,j,k,\rho_i,\rho_j,\rho_k)$. When the sign of the literals is not important, we write simply $C = (i,j,k)$.*

Proposition 13. *A clause $C = (i,j,k,\rho_i,\rho_j,\rho_k)$ excludes a set of 2^{k-3} intervals of length $1/2^k$:*

$$\bigcap_{m\in\{i,j,k\}} \Big(\bigcup_{n=1}^{2^{m-1}} \big| \frac{2(n-1)+\rho_m}{2^m}, \frac{2(n-1)+\rho_m+1}{2^m} \big| \Big)$$

Proof. The exclusion intervals of C are the points that make the literals l_i, l_j, and l_k all false. We can view these intervals as the intersection of the sets of intervals excluded by $\neg l_i$, $\neg l_j$, and $\neg l_k$ respectively. Now, the set of intervals excluded by any literal $\neg l_m$ (with $m \in \{i,j,k\}$) consists of all sequences $(v_l)_{l\in\mathbb{N}^*}$ such that $\nu(v_m) = \rho_m$. Thus, we want to exclude all sequences in $\{0,1\}^{\mathbb{N}^*}$ whose geometric representation is:

$$\phi((v_l)_{l\in\mathbb{N}^*}) = \sum_{l=1}^{m-1} \frac{\nu(v_l)}{2^l} + \frac{\rho_m}{2^m} + \sum_{l=m+1}^{\infty} \frac{\nu(v_l)}{2^l}.$$

Notice that, regardless of the values of $\nu(v_l)$ for $l > m$, the right-hand sum is bounded by

$$0 < \sum_{l=m+1}^{\infty} \frac{\nu(v_i)}{2^l} < \sum_{l=m+1}^{\infty} \frac{1}{2^l} = \frac{1}{2^m}$$

The sequences we wish to exclude are therefore within the following intervals:

$$\bigcup_{(v_l)_{l\in\mathbb{N}^*}} |\sum_{i=1}^{m-1} \frac{\nu(v_i)}{2^i} + \frac{\rho_m}{2^m} + 0, \sum_{i=1}^{m-1} \frac{\nu(v_i)}{2^i} + \frac{\rho_m}{2^m} + \frac{1}{2^m}|$$
$$= \bigcup_{n=1}^{2^{m-1}} |\frac{2(n-1)+\rho_m}{2^m}, \frac{2(n-1)+\rho_m+1}{2^m}|$$

We conclude by taking the intersection of these intervals for $m = i$, $m = j$, and $m = k$. □

Although the expression we have just derived is not particularly easy to manipulate, it exhibits a symmetry that is important to highlight. First, we observe that the width and number of intervals depend only on the index k corresponding to the largest literal l_k. This means that the clauses $\neg(l_3 \wedge \neg l_4 \wedge l_{11})$ and $\neg(\neg l_6 \wedge l_8 \wedge \neg l_{11})$, for example, will both count 2^8 intervals of length 2^{-11}. None of these intervals will overlap, but the gap between two consecutive intervals can only take three possible values: $2/2^i$, $2/2^j$, or $2/2^k$. While k thus controls the number and size of the intervals, the indices i and j define their distributions along the segment $[0, 1]$. As for the signs of the literals, captured by ρ_i, ρ_j, and ρ_k, they shift the intervals by distances of 2^{-i}, 2^{-j}, and 2^{-k} (respectively). In particular, it is worth noting that a clause with at least one positive literal can never exclude the segment $[0, 1/2^k]$. Conversely, a clause with at least one negative literal cannot exclude the segment $[1 - 1/2^k, 1]$. Since there are eight ways to choose ρ_i, ρ_j, and ρ_k, a minimum of eight clauses with the same index k will always be required to exclude the entire segment $[0, 1]$.

Beyond these very general properties, it is interesting to observe that the exclusion intervals follow a fractal evolution. To realise this, it may be helpful to represent the clauses $C_0 = (i, j, k, \rho_1, \rho_2, \rho_3)$ and $C_1 = (i + 1, j + 1, k + 1, \rho_1, \rho_2, \rho_3)$ on a graph. We will use the indicator function $\mathbb{1}_{C_1}$, which takes the value 1 for all real numbers belonging to an exclusion interval, and 0 otherwise. We will quickly identify an interesting symmetry. The graph obtained for C_1 on the segment $[0, 1/2]$ is identical to that of $C_0 = (i, j, k, \rho_1, \rho_2, \rho_3)$ on the entire segment $[0, 1]$, apart from a scaling factor. In fact, it is as though C_1 is duplicated into 2 copies of C_0, which are distributed on either side of the point $1/2$. If we continue the experiment with $C_3 = (i + 3, j + 3, k + 3, \rho_1, \rho_2, \rho_3)$, we will notice exactly the same phenomenon. C_3 indeed presents four duplications of C_0, symmetrically distributed across the four quarters of the segment $[0, 1]$. In general, we can

say that clause C_0 follows the same pattern on the segment $[0,1]$ as clause $C_n = (i+n, j+n, k+n, \rho_1, \rho_2, \rho_3)$ on the segments $[m/2^n, (m+1)/2^n]$ (with $0 < m < 2^n$).

Taking advantage of this periodicity, one might wonder if there is a way to represent the exclusion intervals using a compact formula. The analytical expression of proposition 13 is indeed cumbersome due to its numerous terms. The indicator function $\mathbb{1}_C$ also does not withstand long under calculation. Rather than checking each time whether $x \in [0,1]$ belongs to one of the 2^{k-3} exclusion intervals of clause $C = (i,j,k)$, it would be more convenient to introduce a periodic function.

Définition 26. *Let three positive integers $i < j < k$ and $\rho_i, \rho_j, \rho_k \in \{0,1\}$. The harmonic $\psi = (i,j,k,\rho_i,\rho_j,\rho_k)$ is the function from $[0,1]$ to $\mathbb{R}$:*

$$\psi_{i,j,k}(x) = \min_{m \in \{i,j,k\}} \{\sin(2\pi 2^m x + \rho_m \pi)\},$$

When they exist, we denote ψ^- and ψ^+ as the left and right derivatives of ψ. The number $2/2^k$ is called the wavelength of ψ and is denoted by λ.

The harmonic Ψ is a piecewise sinusoidal function. The behaviour of its negative parts is rather chaotic, but its positive parts exactly trace the exclusion intervals associated with the clause $C = (i,j,k,\rho_i,\rho_j,\rho_k)$. A careful analysis reveals three periods $\lambda_i = 2/2^i, \lambda_j = 2/2^j$, and $\lambda_k = 2/2^k$, which are responsible for the dispersion and symmetries of the harmonic. The "wavelength" of interest is λ_k, as it defines the minimum gap between each exclusion interval, as well as the length of these intervals. For this reason, we will often refer to the wavelength λ, without explicit mention of the index k, whenever the context is sufficiently clear.

Proposition 14. *Let $C = (i,j,k,\rho_i,\rho_j,\rho_k)$ be a clause. Denote by $|a_n, b_n|$ the n-th exclusion interval of C, with $a_1 < a_2 < ... < a_{2^{k-3}} \in \mathbb{R}$ and $b_1 < b_2 < ... < b_{2^{k-3}} \in \mathbb{R}$. For any $x \in [0,1]$, the harmonic $\psi = (i,j,k,\rho_i,\rho_j,\rho_k)$ satisfies:*

- *$\psi(x) > 0$ if and only if there exists an n such that $x \in (a_n, b_n)$;*
- *$\psi(x) = 0$ and $\psi^+ > 0$ if and only if there exists an n such that $x = a_n$;*
- *$\psi(x) = 0$ and $\psi^- < 0$ if and only if there exists an n such that $x = b_n$.*

In the last two cases, we will say that x is a left (respectively right) limit point of ψ.

Proof. Choose $m \in \{i, j, k\}$. According to Proposition 13, the exclusion intervals of the literal $\neg l_m$ are given by:

$$\bigcup_{n=1}^{2^{m-1}} \left| \frac{2(n-1)+\rho_m}{2^m}, \frac{2(n-1)+\rho_m+1}{2^m} \right|$$

It is easy to verify that $\sin(2\pi 2^m x + \rho_m \pi) > 0$ if and only if there exists an $n \geq 1$ such that $2(n-1)\pi < 2\pi 2^m x + \rho_m < 2n\pi$. □

This proposition proves that to every clause C corresponds a unique harmonic ψ. Similarly, to every exclusion interval $|a, b|$ rejecting a part of the Cantor space $\{0, 1\}^{\mathbb{N}^*}$, corresponds an interval (a, b) rejecting a subset of the real line. By abuse of language, we will write indifferently $|a, b|$ or (a, b) to designate an exclusion interval, whenever the context is sufficiently clear.

11.4 Solution Method

To solve 3-SAT$_{\aleph_0}$, we will proceed with an intelligent scan of the segment $[0, 1]$, starting from an initial iterate $x_0 = 0$ and moving to the right whenever a harmonic ψ rejecting this iterate is encountered. If, as we go through the clauses of the problem one by one, none of them rejects x_0, then this value is a solution to the problem, and the method ends. Otherwise, this means that x_0 is a left limit point of a sinusoid ψ with period λ. We can then take $x_1 = x_0 + \lambda/2$ and repeat the procedure.

If several sinusoids $\psi_{i,j,k}$ reject x_0, we must take the one whose period is the largest or — which amounts to the same — the one whose index k is the smallest. By proceeding in this way, we will move the algorithm along more quickly. Moreover, by denoting $\psi_0, \psi_1, ...$ as the sinusoids of indices $k_1, k_2, ...$ rejecting successively the iterates $x_0, x_1, ...$, we ensure that, at every step $\tau \subset \omega_1$ of the algorithm, the iterate $x_\tau = x_{\tau+1} + \lambda_\tau/2$ is always a left limit point of $\psi_\tau(x_\tau)$ and a right limit point of $\psi_{\tau+1}(x_\tau) = 0$. In other words, ψ_τ and $\psi_{\tau+1}$ intersect at x_τ. If this property is unclear to the reader for now, a rigorous proof will follow. For the moment, let us note that, provided this property is correct, the exclusion intervals $|x_0, x_1|, |x_1, x_2|, ..., |x_{\tau-1}, x_\tau|$ traversed by our method after τ iterations can be more simply written as $[0, x_\tau|$.

Algorithme pour 3-SAT$_{\aleph_0}$

Let Φ be the set of harmonics associated with the clauses of the problem and $x = 0$ be the initial iterate of the method.

```
Algorithm (Φ, x)
   λ-max := 0
   foreach (ψ,λ) in Φ
      if (ψ(x) rejects x)
         λ-max := Max(λ, λ-max)
   if λ-max > 0
      x := x + λ/2
      if x = 1
         return "no solution"
      else
         return Algorithm(Φ, x)
   return x
end
```

For efficiency, we could sort the harmonics ψ in descending order of wavelength by issuing an instruction like $trier(\Phi)$ at the start of the algorithm. In this way, we would guarantee that the first harmonic ψ encountered, which excludes the current iterate, is indeed the one with the largest period. This would allow us to immediately exit the *foreach* loop. Note that the descending order ranking, in itself, requires at most $\aleph_0$ steps. Indeed, it involves checking all pairs of constraints. Now, the pairs of constraints can be bijected with $\mathbb{N}^2$ and, using Cantor's pairing function, it is possible to establish a bijection between $\mathbb{N}^2$ and $\mathbb{N}$.

Before showing the convergence of the method, let us note that it could stop at an iterate x_τ breaking the injectivity of ϕ. In this case, the question must be addressed: to which sequence $v \in \{0,1\}^{\mathbb{N}^*}$ does this iterate correspond?

Définition 27. *We say that a real number $x \in [0,1]$ is a solution of 3-SAT$_{\aleph_0}$ on the left (respectively on the right) if it is not the right (respectively left) limit point of any harmonic. A solution is said to be both on the left and on the right if it is not the limit point of any harmonic.*

When x is a solution on the left but not on the right, it means that it is the image by ϕ of a sequence $v \in \{0,1\}^{\mathbb{N}^*}$ that ends with an infinite sequence of 1s. This sequence is on the left boundary of an interval of the form $|\phi(v), b|$ but not on the right boundary of an interval of the form $|a, \phi(v)|$, nor inside any interval. Conversely, if x is a solution on the right but not on the left, the sequence $v \in \{0,1\}^{\mathbb{N}^*}$ whose image is x ends with an infinite sequence of 0s. It lies on the boundary of an exclusion interval of the form $|a, \phi(v)|$.

Proposition 15. *Let C_{i_1,j_1,k_1} and C_{i_2,j_2,k_2} be two constraints with $0 < k_1 \leq k_2$. Let $(I_n)_{n \leq k_1 - 2}$ and $(J_n)_{n \leq k_2 - 2}$ be the collections of disjoint exclusion subintervals associated with these constraints. If there exist $l \leq k_1 - 2$ and $m \leq k_2 - 2$ such that $I_l \cap J_m \neq \emptyset$, then $J_m \subseteq I_l$.*

Proof. By proposition 13, the subintervals I_l and J_m are written as

$$I_l = |\frac{(n-1)}{2^{k_1}}, \frac{n}{2^{k_1}}| \qquad\qquad J_m = |\frac{(n'-1)}{2^{k_2}}, \frac{n'}{2^{k_2}}|$$

for positive integers n and n' that are non-zero. Suppose that $I_l \cap J_m \neq \emptyset$ but that J_m is not a subset of I_l. It must then be the case that $(n'-1)/2^{k_2} < n/2^{k_1} < n'/2^{k_2}$ or $(n'-1)/2^{k_2} < (n-1)/2^{k_1} < n'/2^{k_2}$. But since $k_1 \leq k_2$, there exists a positive integer p such that $2^{k_1} = 2^{k_2}/p$. Substituting into the relations above and simplifying, it must be either that $(n'-1) < np < n'$, or $(n'-1) < (n-1)p < n'$. Neither of these relations can be satisfied because np and $(n-1)p$ are positive integers. We conclude that $J_m \subseteq I_l$. □

Proposition 16. *Let $x_\tau \in [0,1]$ be the point obtained after $\tau \subset \omega_1$ iterations of the algorithm. Suppose that the solution is not in $[0, x_\tau|$ and that all the harmonics ψ that reject x_τ on the left have this point as their left limit. In this case, exactly one of the following two propositions is correct:*

1. *x_τ is a solution of the problem on the left;*
2. *At the next step of the algorithm, the solution of the problem is not in $[0, x_{\tau+1}|$ and all exclusion harmonics ψ that reject $x_{\tau+1}$ on the left have this point as their left limit.*

Proof. Consider the family $(\psi)_i$ of exclusion harmonics rejecting x_τ on the left. If this family is null, then point 1 is true. Otherwise, let ψ_τ be the one of these harmonics with the largest wavelength λ_τ. By assumption, every harmonic $\psi_j \in (\psi)_i$ has x_τ as its left limit point and thus admits the exclusion subinterval $I_j = (x_\tau, x_\tau + \lambda_j/2)$ (where λ_j is the wavelength of ψ_j). By the continuity of ψ_j and ψ_τ, there exists a real number $\epsilon_j > 0$ such that $x_\tau + \epsilon_j \in (I_j \cap I_{x_\tau})$. By the previous proposition, we thus have $I_j \subseteq I_\tau$. This proves that the solution is not in $[x_\tau, x_\tau + \lambda_\tau/2|$. Now, by assumption, we also know that the solution is not in $[0, x_\tau|$. The next iterate $x_{\tau+1} = x_\tau + \lambda_\tau/2$ given by the algorithm is such that $[0, x_\tau| \cup [x_\tau, x_\tau + \lambda_\tau/2| = [0, x_{\tau+1}|$ (proposition 12). This proves the premise of point 2.

Suppose now that there exists a harmonic ψ_j rejecting $x_{\tau+1}$ on the left and not having $x_{\tau+1}$ as its left limit point (it thus also rejects $x_{\tau+1}$ on the right). Let I_j be the exclusion subinterval of ψ_j containing $x_{\tau+1}$. By the continuity of ψ_j and ψ_τ, there exists a real number $\epsilon_j > 0$ such that $x_{\tau+1} - \epsilon_j \in (I_j \cap I_{x_\tau})$. By the previous proposition, it must then either be that $I_j \subseteq I_\tau$ (if $\lambda_j < \lambda_\tau$), or that $I_\tau \subseteq I_j$ (if $\lambda_j > \lambda_\tau$). In the first case, this

contradicts the fact that ψ_j rejects $x_{\tau+1}$ on the left, because $x_{\tau+1}$ is a right limit point of $I_\tau = (x_\tau, x_\tau + \lambda_\tau/2)$. In the second case, it is incompatible with the construction of ψ_τ. Without loss of generality, either I_j does not have x_τ as its left limit point, or it does. If it does not, then $I_\tau \subseteq I_j$ does not have it as well, which contradicts the fact that ψ_τ belongs to the family $(\psi)_i$ defined at the beginning of the proof. But if I_j has x_τ as its left limit point, then ψ_j also belongs to $(\psi)_i$ and λ_τ is no longer the harmonic with the largest wavelength in this family (since $\lambda_\tau < \lambda_j$). □

Geometrically, the proposition above amounts to saying that the functions $\psi_0, \psi_1, \ldots, \psi_\tau$ constructed by the method have the iterates $x_0, x_1, \ldots, x_\tau$ as their left-hand limits. Step by step, these functions construct an exclusion interval $[x_0, x_1| \cap |x_1, x_2| \cap |x_{\tau-1}, x_\tau| = [x_0, x_\tau|$ that grows larger, until it reaches 1 or finds a solution. However, one must still ask what happens when the wavelengths $\lambda_1, \lambda_2, \ldots$ of the functions $\psi_1, \psi_2, \ldots$ encountered by the method at each iteration tend towards 0? Is the point x obtained at the end of this limit a solution? Otherwise, we have no reason, *a priori*, to think that such a situation could not occur many times, to the point, perhaps, of preventing the method from converging in a countable time.

Proposition 17. *After $\tau \subset \omega_1$ iterations, suppose that the algorithm produces an infinite sequence of iterates $(x_{\tau+n})_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} x_{\tau+n} = x$. If the solution is not in $[0, x_\tau|$ and all exclusion harmonics ψ that reject x_τ to the left have this point as their left-hand limit, then x is a solution to the problem on the right.*

Proof. For every $n \in \mathbb{N}$, consider the following two properties: (i) the solution is not in $[0, x_{\tau+n}|$ and (ii) all exclusion harmonics ψ that reject $x_{\tau+n}$ to the left have this point as their left-hand limit. By assumption, (i) and (ii) are true for $n = 0$. Proposition 16 tells us that if these properties hold for $x_{\tau+n}$, then they also hold for $x_{\tau+n+1}$. By induction, we conclude that for every $n \in \mathbb{N}$, every $x_{\tau+n}$ satisfies properties (i) and (ii).

Now suppose, by contradiction, that there exists a harmonic ψ that rejects x to the right. By the continuity of ψ, there exists an $\epsilon > 0$ such that every real $y \in (x - \epsilon, x)$ is rejected by ψ both to the right and to the left. Since $\lim_{n\to\infty} x_{\tau+n} = x$, there exists an $N \in \mathbb{N}$ such that $x_{\tau+N} \in (x - \epsilon, x)$. This point $x_{\tau+N}$ is therefore rejected by ψ both to the left and to the right. In other words, ψ rejects this point to the left without having it as a left-hand limit, which contradicts property (ii). We conclude that x is indeed a solution to the problem on the right. □

We are now in a position to iteratively prove the validity of the algorithm. At the first step, the iterate $x_0 \in [0, 1]$ is zero. If no constraint rejects it, then it is a solution to the problem. Otherwise, the hypothesis of Proposition 16 holds, as 0 is a left-hand limit point of all the intervals rejecting 0. By

induction, and using Proposition 16, we conclude that at each step τ of the algorithm, either x_τ is a solution to the problem, or it lies on the left-hand edge of an exclusion harmonic ψ_τ of wavelength λ_τ. In the latter case, we can then restart the algorithm by setting $x_{\tau+1} = x_\tau + \lambda_\tau/2$. In the special case where $x_{\tau+1} = 1$, Proposition 16 tells us that the solution is not in $[0, x_\tau + \lambda_\tau| = [0, 1| = [0, 1]$ (the last equality follows from Proposition 12). It is therefore easy to see that x_τ will increase at each step of the algorithm, until it reaches 1 or halts on a solution.

It remains to show that convergence occurs, at most, after $\mathcal{O}(\omega_0)$ iterations. To do this, let us first note that at each call to the *while* loop of the algorithm, the latter performs at most a number of operations proportional to the number of clauses, which is itself countably infinite. Moreover, we can establish a bijection between each call of the *while* loop and the iterate x_τ computed during that call. As long as x_τ is rejected by a clause, x_τ is a rational number. Indeed, we start with $x_0 = 0$ (a rational number) and add, at each iteration, the number $\lambda_\tau/2 = 1/2^k$ (for some $k \in \mathbb{N}$), which is also a rational number. Furthermore, there is no $\tau' < \tau$ such that $x_\tau = x_{\tau'}$. If this were the case, we would violate the construction rule of the algorithm, which sets $x_{\tau+1} = x_\tau + \lambda_\tau/2 > x_\tau$. Therefore, the τ-th call of the *while* loop, if it does not converge to a solution, can be associated with a distinct rational number x_τ. In other words, there is an injection between the set of algorithm steps and the set of rational numbers. Since $\mathbb{Q}$ has cardinality $\aleph_0$, we conclude that the algorithm has complexity at most $\mathcal{O}(\omega_0)$. This reasoning is summarised in the following theorem.

Solubility of 3-SAT$_{\aleph_0}$

The 3-SAT$_{\aleph_0}$ problem is solvable, in the worst case, in at most a countably infinite number of steps. In the particular case where the problem has no solution, it is solvable in finite time.

Let us note that if the clauses have been sorted in descending order of wavelength, a 3-SAT$_{\aleph_0}$ problem with no solution can be completed in a finite number of steps. Indeed, each iteration of the *while* loop will take a finite time. Moreover, there can only be a finite number of calls to this loop; otherwise, it would mean that the iterates $v_0, v_1, ...$ would form an infinite series. However, proposition 17 shows that such a series always converges to a solution.

If the clauses are not presented in increasing order of wavelengths, the program could still decide to retain by default the first harmonic ψ, rejecting the iterate x_τ it encounters. Let (a, b) denote the exclusion interval of ψ in which (or on the boundary of which) x_τ lies at the τ-th iteration. Instead of setting $x_{\tau+1} = x_\tau + \lambda$, the algorithm would set $x_\tau = b$. In these conditions,

it would also converge to a solution, albeit less "efficiently" (the notion of efficiency is, however, relative in the context of transfinite machines, due to the equipotency of countable ordinals). However, the proof requires that the clauses be of ordinality ω_0. If this is not the case, we know, by proposition 8, that it is possible to simulate a machine with inputs of size greater than ω_0 by limiting it to a tape of size ω_0. This simply requires presenting the 3-SAT$_{\aleph_0}$ problem in an appropriate form, without constituting a restriction of a properly conceptual order.

Proposition 18. *The impossibility of the 3-SAT$_{\aleph_0}$ problem, with exactly ω_0 clauses and ω_0 variables, can be determined in finite time.*

Proof. The proof follows directly from what has just been stated. □

We conclude that a finite program can determine, in finite time, whether the 3-SAT$_{\aleph_0}$ problem has a solution. A fortiori, this also implies that it can determine whether a transfinite machine running this program halts in finite time. Thus, we have exhibited an example of a finite program capable of determining whether an infinite Turing machine $\mathbf{T}_1$ halts. This machine would be the restriction of $\mathbf{T}_1$ to a finite number of clauses and propositional variables, whose behaviour would be identical to that of $\mathbf{T}_1$ in finite time. In other words, this means that an impossible 3-SAT$_{\aleph_0}$ problem contains no more information than a certain judiciously chosen 3-SAT problem. Or again, there exists a finite number of clauses which, taken together, exclude all admissible solutions. The additional clauses and variables represent, at best, an unnecessary redundancy.

One question one might ask is whether an impossible 3-SAT$_{\aleph_0}$ problem, from which the impossible 3-SAT subproblem has been extracted, would become solvable or not? This requires precisely defining what we mean by "extraction". It would probably involve removing the subset of mutually exclusive clauses traversed by the finite machine to determine the impossibility of 3-SAT$_{\aleph_0}$, as well as all the propositional variables that, following this extraction, would remain free to take any value. If this process were done iteratively, would it lead to a solvable 3-SAT$_{\aleph_0}$ problem? If the number of extractions is finite, this goes without saying, since the finite union of finite sets of clauses results in a finite number of clauses. However, 3-SAT$_{\aleph_0}$ is composed of a countably infinite number of clauses. On the other hand, it could be that by performing an infinite number of extractions, one ends up removing the entire set of clauses or reducing 3-SAT$_{\aleph_0}$ to a finite and solvable 3-SAT problem. This infinite number of extractions cannot be performed by a finite program, but by a machine $\mathbf{T}_1$. Therefore, one could write an infinite program that determines whether a 3-SAT$_{\aleph_0}$ problem is solvable or not in infinite time. This line of reasoning represents an alternative proof, and a confirmation, of the solvability of 3-SAT$_{\aleph_0}$ by a transfinite

program. What we would now like to do is show that the concept of the transfinite machine $\mathbf{T}_1$ that we introduced at the beginning of the document can implement these kinds of infinite programs, of which we have, until now, only given the pseudocode.

Proposition 19. *The 3-SAT$_{\aleph_0}$ problem is solvable by a transfinite Turing machine* $\mathbf{T}_1 = (K, \Sigma, \delta, s, H, \alpha)$.

Proof. We know that there exists an encoding that allows the clauses and variables of the problem to be expressed using a string $x \in \Sigma^*$, as is the case for a classical Turing machine. We can assume, without loss of generality, that x is of size equal to ω_0. Proposition 8 indeed states that a Turing machine using a tape of larger size can be simulated by a machine using a tape of length ω_0. In these conditions, proposition 18 teaches us that, if the program has no solutions, there exists a method that determines this fact in finite time $\tau \subset \omega_1$. By proposition 5, a transfinite machine $\mathbf{T}_1$ that halts in finite time behaves like its finite restriction. However, implementing the algorithm 11.4 on a finite machine presents no difficulty. We can deduce that our machine $\mathbf{T}_1$ will only reach $|_0\triangleright$ if the 3-SAT$_{\aleph_0}$ problem has a solution. Therefore, it is sufficient to instruct the machine to halt as soon as it reaches the symbol $|_0\triangleright$, whether in the state ∞ or in any other. According to proposition 17, the encoding of the iterate x_{ω_0}, present on the tape at that moment, is the solution to the problem. □

11.5 Program for a Transfinite Machine

Although the harmonics ψ are elegant for understanding the method and proving its convergence, they are not the most convenient for computational calculations. It is more difficult to calculate a sinusoid than to verify a binary constraint of the form $l_i \vee l_j \vee l_k$. Furthermore, the calculation of the sine may not be exact if truncated before the ω_0-th decimal place, meaning that it would take up to ω_0 steps to verify a 3-SAT$_{\aleph_0}$ clause, whereas this could be done in just three steps.

To avoid handling sinusoids, the pseudocode will, at each step $\tau \subset \omega_1$, construct a vector v_τ with values in $\{0,1\}^{\mathbb{N}}$, corresponding to a solution of 3-SAT$_{\aleph_0}$. The program checks a clause $l_i \vee l_j \vee l_k$ by setting $\nu(v_i)$, $\nu(v_j)$, and $\nu(v_k)$ equal to the i-th, j-th, and k-th coordinates of the vector v_τ in progress. The clauses themselves are represented in a matrix C of dimension $\aleph_0 \times 6$, which is given as input to the algorithm. Each row of this matrix is written in the form $(n_1, n_2, n_3, a_1, a_2, a_3)$. The integers $n_1, n_2, n_3 \in \mathbb{N}^*$ refer to the indices of three propositional variables $v_{n_1}, v_{n_2}, v_{n_3}$, and the binary numbers $a_1, a_2, a_3 \in \{0,1\}^{\mathbb{N}^*}$ denote the polarity of the literals (positive or negative), following the same convention as that used in Proposition 14. For

example, the row $(2, 4, 6, 0, 1, 0)$ represents the constraint $\neg(\neg v_2 \wedge v_4 \wedge \neg v_6)$[2].

Each time the program encounters a constraint at index k that excludes the solution, it will perform a binary addition $v_{\tau+1} = v_\tau + v_k$, where $v_k \in \{0,1\}^{\mathbb{N}^*}$ denotes a vector whose all coordinates are zero except for the k-th one. For example, if $v_\tau = 01011000...$ and $v_k = 00001000...$, then the binary addition yields $01100000....$ In this way, we have $\phi(v_{\tau+1}) = \phi(v_\tau) + 1/2^k$. In the particular case where the i-th coordinate of the vector v_τ is equal to 1 for all $i \geq k$, the addition is impossible, and the algorithm terminates. Geometrically, this corresponds to a situation where $\phi(v_\tau) + 1/2^k \geq 1$. We can therefore conclude that there is no solution.

This situation may arise, for example, if the algorithm has already excluded the interval $[0, 15/16]$ and is investigating whether a solution exists within the segment $[15/16, 1]$. At this point, the vector v_τ is $111100\ldots$. In geometric terms, this vector corresponds to a position on the real line equal to $1/2 + 1/4 + 1/8 + 1/16 = 15/16$, which we write as $\phi(v_\tau) = 15/16$. Suppose that, in the next step of the algorithm, the constraint $\neg(v_1 \wedge v_2 \wedge v_3)$ is encountered. This constraint excludes the interval $[7/8, 1]$, thereby leaving no possible solutions.

From the perspective of the algorithm, the discovery of a new exclusion interval means moving along the real line by a distance equal to the length of the newly excluded interval; in this case, $1/8$. Technically, this requires performing the operation $v_{\tau+1} = v_\tau + v_k$, with $v_\tau = 111100\ldots$ and $v_k = 00100000\ldots$. The binary addition of these vectors is equivalent to $\phi(v_\tau) + \phi(v_k) = 15/16 + 1/8$, which yields a value greater than 1. However, there exists no vector v in Cantor space such that $\phi(v) > 1$. Thus, the addition is impossible, and the algorithm terminates.

```
Algorithm (C)
   v := vectorZero();
   sortByAscendingOrder(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
```

[2]For the sake of clarity, let us note that the matrix C contains as many rows as there are constraints in the generalised 3-SAT problem; that is, a countably infinite number. Whenever we wish to evaluate the constraint C_i, it is necessary to combine the information in the ith row of the matrix with the values of the vector v_τ. This vector evolves over time. If a solution exists, the vector must converge to a solution in, at most, a countably infinite time. This would not be possible with a brute-force algorithm, which tries every possible value of τ, yet by reading the constraints intelligently, I have previously demonstrated that it is possible to force τ to converge within countably infinite time (if a solution exists), or to reach a finite-time conclusion that no solution exists.

```
            n := n3
            if(v(n) != 1)
               v(n) := 1
            else if (n > 1)
               n := n-1
            else
               return "no solution"
            end
         end
      end
   end
   return v
end
```

This program corroborates the proof that a transfinite Turing machine $\mathbf{T}_1$ can solve a generalized NP-complete problem in a countably infinite time $\aleph_0$, and not in an uncountably infinite time $\aleph_0^2$. This means that it is just as easy for $\mathbf{T}_1$ to find a solution as it is to verify a solution. Or, to put it differently, one can say that the class of transfinite problems P^T is equal to the class of problems NP^T, which is denoted $P^T = NP^T$.

The first question that remains open is whether transfinite NP-complete problems correspond to real-world problems that our minds are regularly exposed to. Indeed, I have shown that transfinite NP-complete problems can be represented on the real line; that they are problems defined in the continuum. However, it is not certain that these problems are not just a simple fabrication, a mathematical artefact, created specifically for the needs of my argument—one with no practical application in everyday human life.

The second important question is whether our mind actually functions like a transfinite Turing machine. If, indeed, generalized NP-complete problems are tasks that the human mind is generally tasked with solving, we do not understand why this task could not be overcome with, for example, a neural network. The hypothesis that the mind functions like a Turing machine is concerning for the simple reason that our brain, in contrast, is finite in size. It is precisely for this reason that I distinguished the notion of a model from that of a function. I have constantly insisted that the Turing machine is a model of intuition, not of the mind per se. It may therefore be the case that our mind does not function at all like a transfinite Turing machine; however, its operation must still support the same types of functions that the transfinite Turing machine is capable of.

I will address these delicate questions in the next chapter. I will also take the opportunity to sketch a coherent philosophical view on the relationship between consciousness and the brain. But before that, I would like to touch upon our central result, namely that $P^T = NP^T$, and wonder whether this might teach us something relevant about the problem $P? = NP$. At first

glance, it seems that a result obtained on transfinite NP-complete problems could be transposed, at least partially, to the case of finite NP-complete problems. I will show, however, that nothing is less certain; I will, however, outline a list of promising strategies.

11.6 Strategies Concerning $P = NP$

The mathematician Charles Fefferman is said to have once remarked: "Writing a proof is like playing chess with the Devil. Although you lose continuously, you have the right to go back as many times as necessary." Seen from this perspective, studying a mathematical problem as tough as $\mathcal{P} = \mathcal{NP}$ may be akin to analysing a well-documented chessboard position. We know that the position of the pieces can ultimately lead to only three possible outcomes: victory ($\mathcal{P} = \mathcal{NP}$), defeat ($\mathcal{P} \neq \mathcal{NP}$), or a draw ($\mathcal{P} = \mathcal{NP}$ is undecidable), yet no decisive argument seems to be forthcoming. Fortunately, there is a vast difference between checkmate and pure mathematical proof. While the former is achieved by reducing the complexity of a particular position through well-thought-out moves (and perhaps this is what a certain branch of algorithmic science seeks to emulate), the latter, by contrast, proceeds through ever-widening generalisations (even at the cost of losing sight of concrete objects).

It is unlikely that a seasoned chess player would ever consider transposing a difficult opening onto a $\aleph_0 \times \aleph_0$ chessboard? As strange as this may sound, I believe that, with respect to the search for a mathematical proof, formulating an NP-complete problem using an infinite set of variables and constraints might provide some interesting insights.

This is precisely what I did with the 3-SAT problem. This problem is NP-complete for a classical Turing machine. However, I generalised it to a countably infinite set of propositional variables $D = \{v_1, v_2, ...\}$ and clauses $l_i \vee l_j \vee l_k$ (with i, j, k distinct natural numbers), where the literal l_i denotes either the variable v_i (a positive literal) or $\neg v_i$ (a negative literal). Any admissible solution to this new problem 3-SAT$_{\aleph_0}$ consists in assigning the value *true* to a subset of propositional variables $V \subseteq D$ and the value $false$ to its complement $V^{\mathcal{C}}$. The space of possible solutions, for its part, is given by the set of subsets of $\mathbb{P}(D)$. In this context, I constructed a transfinite Turing machine $\mathbf{T}_{\omega_1}$. This machine behaves, in all respects, like a classical Turing machine, except that it is capable of performing a countably infinite number of successive operations (formally, we would say that the number of operations α performed must satisfy $\alpha < \omega_1$).

It is easy to see that our machine $\mathbf{T}_{\omega_1}$ can verify a solution to the problem 3-SAT$_{\aleph_0}$ in a countably infinite amount of time. More specifically, we would say that the number of steps required can be put into bijection with the $\aleph_0$ clauses defining the problem. On the other hand, to go through the

set of possible solutions one by one, we would need to consider all elements of $\mathbb{P}(D)$, whose cardinality is uncountable. Unless "luck" is on our side, or unless we are faced with a problem admitting a "large number of solutions" (in a sense that we will clarify), a machine $\mathbf{T}_{\omega_1}$ would not be able to solve the 3-SAT$_{\aleph_0}$ problem by brute force. This situation is not unrelated to the classical 3-SAT problem. It is therefore natural to wonder if there is any connexion between the following two questions:

Q1 : Is there a deterministic Turing machine capable of solving any instance of 3-SAT in polynomial time (in the worst case)?

Q2 : Is there a deterministic *transfinite* Turing machine capable of solving any instance of the 3-SAT$_{\aleph_0}$ problem in a countably infinite number of operations (in the worst case)?

The developments I presented in the previous sections shed light on this question. On the one hand, they provide an affirmative answer to question Q2. On the other hand, and this is the most difficult point, they may allow us to better understand the relationship between questions Q1 and Q2 (assuming there is only one, for it is not so obvious). Several possibilities are conceivable. One might think that if we had an algorithm capable of solving the 3-SAT problem in polynomial time on a classical Turing machine, then we could generalise this algorithm on a transfinite Turing machine in such a way as to solve the 3-SAT$_{\aleph_0}$ problem. This would mean that Q1 $\rightarrow$ Q2. By contraposition, one could then show $\mathcal{P} \neq \mathcal{NP}$ through a simple refutation of Q2.

It could also be suggested, in contrast to this first approach, that a positive answer to question Q2 perhaps implies a positive answer to question Q1. Once one is capable of constructing an algorithm solving the 3-SAT$_{\aleph_0}$ problem, it would be interesting to transpose the principles of this algorithm onto a classical Turing machine, hoping thus to solve the 3-SAT problem. This would lead to a constructive proof that $\mathcal{P} = \mathcal{NP}$. Even in the absence of a constructive proof, a simple general result showing that Q2 true implies Q1 true would be enough to resolve the problem, provided it is established, in one way or another, that Q2 is true.

The intuitively less probable situations, where Q1 true implies Q2 false and Q2 true implies Q1 false, are just as interesting as the two situations outlined above. However, it is difficult to conceive how the solution to the 3-SAT$_{\aleph_0}$ problem could lead to the conclusion that a polynomial-time algorithm for the classical 3-SAT problem is impossible, unless one manages to highlight an essential difference between the two problems, explaining why what works in the infinite case is precisely what prevents an algorithm from

	Strategies	Q2 true	Q2 false	Q2 undecidable
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	$\nvdash$ Q2 $\rightarrow \nvdash$ Q1	-	-	undecidable
S6	$\nvdash$ Q1 $\rightarrow \nvdash$ Q2	decidable	decidable	-
S7 ?!	$\vdash$ Q2 $\rightarrow \nvdash$ Q1	undecidable	undecidable	-
S8 ?!	$\nvdash$ Q2 $\rightarrow \vdash$ Q1	-	-	decidable

Table 11.1: This table summarises the main strategies to consider, with the exception of those where undecidability/decidability would imply truth/-falsehood (and vice versa). The strategies marked with the symbol "?!" are considered dubious. By convention, the symbols $\nvdash$ and $\vdash$ denote undecidability (respectively, provability).

working in the finite case. Given the sometimes unpredictable behaviour of infinite sets, this path is not to be excluded *a priori* (though we consider it less reliable).

We can also study what the undecidability of Q2 would imply for the truth of Q1 — and vice versa. For this reason, a rather natural question would be to ask whether the halting problem is solvable, or not, for transfinite Turing machines, and under what conditions. If we had a universal transfinite Turing machine, denoted $\mathcal{U}_{\omega_1}$, capable of determining whether another transfinite Turing machine halts or not, then perhaps we could apply $\mathcal{U}_{\omega_1}$ to the finite input x corresponding to the program of a finite Turing machine, and see if we can obtain the answer in finite time. In this way, we would have solved the halting problem for classical Turing machines, which would lead to a contradiction. This kind of strategy might prove fruitful for (a) showing that the halting problem also exists for transfinite machines and (b) attempting to find a relationship between the decidability of Q1 and Q2.

The set of these strategies is summarised in the table opposite. I consider myself fortunate if, through a careful analysis of the problem, I have been able to exclude certain approaches and keep others open. It should be noted that the answer to question Q2 may depend, at least partially, on the precise definition given to a transfinite Turing machine. Therefore, we do not rule out the possibility of constructing variants of transfinite machines (with potentially pathological behaviours) for which Q2 may not hold the same truth value as the one I have used in this thesis. It is possible that one of these variants could lead to more relevant results. However, I have attempted to define my concept in the most general and operational manner that was permitted to me. Further research is warranted on these important

issues.

What to take away from this chapter

One of the three non-computable functions of human intuition consists in having "luck" when faced with an NP-complete problem, thus solving it quickly.

In this chapter, I have shown that a transfinite Turing machine consistently has "luck" when faced with transfinite NP-complete problems. This means that, in the transfinite case, solving an NP-complete problem is no more difficult than verifying it.

I have also provided a geometric interpretation of the transfinite NP-complete problem, which consists in finding the zeros of a function on the real line. It follows that a transfinite Turing machine can be seen as a machine capable of thinking in the continuous (and thus calculate second order logic and second order arithmetic).

However, it is important to understand that this transfinite Turing machine is not a model of the human mind, but rather a model of *human intuition* (the intuition of the whole and the intuition of the infinite); just as the computer was not a model of the human mind, but of *human reason* (pure thought).

In the next part, I would like to ask what the transfinite Turing machine can nevertheless teach us about the nature of the human mind. It remains difficult to understand how a finite brain is capable of performing functions that belong to a transfinite Turing machine.

Towards a General Theory of the Human Mind

What Neuroscience Tells Us

Chapter 12

The Mind-Body Problem: Perspectivist Mysterianism

In this brief detour into the realm of transfinite Turing machines, we have encountered a significant problem. On the one hand, I have argued that intuition can be explicitly conceived in the form of three non-computable functions: (1) a function for completing infinite series, (2) a function for solving undecidable problems, and (3) a "chance" function that helps find a good initial iterate within NP-complete problems. On the other hand, I have provided a mathematical model supporting these three functions jointly: transfinite Turing machines.

However, I have acknowledged that the transfinite Turing machine, although it models these three functions of intuition, is not itself a model of the human mind. There is, indeed, a significant difference between "modeling functions of intuition" and "modeling the human mind." In the former case, one provides a machine capable of reproducing certain key results of intuition, such as the grasping of infinity. In the latter case, one provides a machine capable of simulating all the dimensions of thought, emotionality, and human sensation. The difference between these two machines is of the same order as the difference between a calculator (a model of arithmetic reasoning) and the human brain.

It may thus be that the model of the human mind, if it exists, is ultimately nothing like the transfinite Turing machines. The question that arises, therefore, is to understand what, if anything, these transfinite machines can nevertheless teach us about the nature of human cognition. There are two ways to answer, both equally problematic in their own way.

1. **Optimistic Answer:** The human mind can be represented as a program implementable by a transfinite Turing machine.

2. **Pessimistic Answer:** The human mind, if it is modelable, cannot be implemented by a transfinite Turing machine.

If we opt for the optimistic answer, asserting that the model of the mind is a certain program of the transfinite Turing machine, that is, a succession of biophysical laws, properties, and relations that can be modeled using the transfinite Turing machine, we encounter a physical impossibility. On what grounds could a finite physical system (our brain, our organism...) perform transfinite operations? This problem is equivalent to the one that would arise from attempting to build a concrete mechanical version of the transfinite Turing machine, a sort of supercomputer.

One way to circumvent this problem is to argue that the transfinite Turing machine is indeed a model of the human mind, but an idealized model. It is, in fact, common in science to use infinite models to represent finite phenomena. A good example is Euclidean theory, which assumes that objects in space, such as straight lines, are infinitely extended — even though in practice, our universe is finite in size. The fact that the model presupposes infinity does not seem to bother us excessively. We can apply the infinite model to finite cases. The ideality of the model serves as a horizon for practical thought. It allows us to define and manipulate geometric objects more easily.

However, this response, which treats transfinite machines as an idealized model, is not entirely convincing. If the transfinite Turing machines constitute an idealized model of the human mind, but in practice the human mind only performs a finite number of steps, would it not ultimately mean that the model of the human mind is a classical Turing machine? Indeed, we have seen that a transfinite Turing machine that performs a finite number of steps behaves in exactly the same way as a classical Turing machine (see proposition 5).

Therefore, one might be tempted to say — and this is the pessimistic answer — that the human mind is not modeled by a transfinite Turing machine, but by a *finite* machine functioning similarly to a transfinite Turing machine. Is it not counterintuitive to suggest that a finite machine could behave like a transfinite machine? One conceivable example would be a classical Turing machine equipped with an oracle. As a reminder, an oracle is a device, with a mysterious functioning, that can provide an answer to a predefined question. In the case that concerns us, the question could be: "What is the convergence of this infinite series?" Our machine would behave exactly like a classical Turing machine, but it could, from time to time, provide an answer to this particular question, bringing it significantly closer to the functioning of a transfinite Turing machine (I have more specifically described the relationship between the transfinite Turing machine and oracles in section 10.3.1).

According to the pessimistic answer, the model of the mind would thus be a classical Turing machine equipped with a kind of black box capable of answering questions about infinity. We do not know how this box works internally; all we can say is (1) that it is not a classical Turing machine and (2) that it associates the correct answer with each question. This amounts to saying that we cannot explicate how the mind functions; that there is an impenetrable mystery within it. Our pessimistic answer thus leads us back to the dilemma of explication that I have described in chapter 9. Ultimately, it amounts to saying that there exists an element X (the oracle), with such and such properties that we cannot explain.

Of course, one could try to overcome the mystery of element X by suggesting that the oracle, although inexplicable within the formalism of a classical Turing machine, could be modeled by a non-classical physical system, perhaps following the laws of quantum mechanics. It is entirely possible that further study of the human brain may one day reveal physical behaviors within cells that do not conform to classical physics. If it is demonstrated that these behaviors (1) cannot be modeled by a classical Turing machine and (2) satisfy at least one of the three non-computable functions of intuition, we would possess an interesting key to understanding the functioning of the black box of intuition.

This approach I have just described, it should be noted, is not without recalling Roger Penrose's strategy. In his book *The Emperor's New Mind*, Penrose attributes certain properties of consciousness to the quantum functioning of the brain. He even seeks to describe a mechanism as precisely as possible. It would go beyond the scope of this thesis to discuss the technical aspects of his approach. However, there are three very general reasons why I do not wish to follow Penrose's lead. I will present them below.

The first reason I wish to avoid resorting to quantum physics to explain the black box of intuition is that it is dangerous, within the framework of speculative philosophy, to base conclusions on the empirical functioning of the brain. The strength of philosophy lies in discussing the necessary relationships between concepts, not contingent physiological mechanisms. Given the complexity of our cerebral organ, to which even neuroscience experts still acknowledge many mysteries, it is unlikely that philosophy, starting from universal ideas, could arrive at precise conclusions.

On the other hand, and this is the second reason, one could easily substitute one mystery for another. Saying that this oracle, this enigmatic black box, is the product of quantum phenomena, without giving these phenomena the full validity of a scientific theory, is to move away from speculative philosophy and venture, very awkwardly, into the realm of science. Far from helping to legitimise philosophy, this would push us to recognise that we have not particularly advanced in elucidating the mystery that we sought to reveal in the full light of day. As Maurice Maeterlinck, Nobel Prize winner in literature, said: "It is rare that a mystery disappears; it usually just changes

its place."

Finally, a third, much simpler reason why I do not wish to resort to quantum physics is this: argumentative rhetoric benefits from always considering the most pessimistic case. Rather than venturing into shaky explanations of the explicit functioning of our black box, let us accept with my critics — those readers we always imagine to be fond of perverse objections — that this black box of intuition cannot receive an ultimate elucidation of its internal workings. Let us ask ourselves, under this hypothesis, whether it invalidates the conceptions I have defended so far — and to what extent.

12.1 The Unsustainable Unity of Body and Mind

I would like to take seriously the seemingly paradoxical premise that the model of the mind is given by a transfinite Turing machine, while the model of the brain is given by a finite mechanism. A priori, these two statements are contradictory. They are only so, however, if we admit that the model of the mind is the model of the brain, and vice versa. Why should the way our mind works be reducible, in its most fundamental terms, to our understanding of the brain? What other alternative is there to this philosophical stance? How can we think about the relationship between the mind (an infinite model) and the brain (a finite model)?

In this chapter, I would like to defend a perspectivist form of mysterianism. To recap, mysterianism asserts that there are mysteries in nature that the human mind is incapable of elucidating. We have seen that philosopher McGinn, for example, maintains that consciousness is part of this mystery (see section 9.4.2). I would like to follow in his footsteps, with a few distinctions. I do not claim that the human mind, in itself, is inexplicable. What is inexplicable, in my view, is the link between the model of the mind and the model of the human brain.

Let us clarify this idea. First, I assert that it is possible (1) to model the mind using physiological theories describing the functioning of the brain's cells; this is the aim of neuroscience. But I also assert that it is possible (2) to model the mind using an introspective psychological theory. This psychological theory does not aim to explain the functioning of the brain through black boxes or quantum laws, but to account for the internal dynamics of the psyche from a subjective — so to speak, phenomenological — perspective. Examples of this approach can be found in the philosophy of Husserl, who studies the principles by which objects are constituted in consciousness. We can also think of Kantian philosophy, which is not fundamentally distant from a structural theory of the human mind. Not to mention Freud's opographic Model, which attempts to give meaning to concepts such as "Ego", "drive", "anxiety", or "guilt". These concepts do not have clearly identifiable physiological substrates. The basic variables are

not biochemical reactions, but lived experiences of consciousness: desires, representations, ideas, thoughts, emotions, etc[1].

What remains mysterious, even in the presence of these (1) physiological and (2) psychological models, is not the human mind per se, but the relationship between our mind and our body — and more generally, the brain. It is well understood that the psyche has laws, drives, and dynamics; life experience and introspection allow us to glimpse them. What we scarcely understand, however, is the relationship between our brain and the lived experiences of consciousness (the sensation of cold, desire, the death...). Is there a place of causality between our thoughts and the functioning of our nerve cells? Is this link unidirectional or bidirectional? How can we explain that something as immaterial as thought is caused by changes within matter?

The perception lived in the first person does not seem to be of the same order as the observation of our brain in the third person. This is why there is an incurable feeling of strangeness when our eyes, by chance, lose themselves in the contemplation of our arms, our head, our lips troubled by nature's obtuse mysteries. It would seem that the feeling of being alive does not coincide with the representation, entirely external and objective, of this body that we are.

The mysterianism I will defend in this chapter will consist of thinking that, if the relationship between body and mind is enigmatic, it is because it is a single reality — the fundamental unity of which we are unable to comprehend. We are unable to do so because of the very nature of our intelligence, and not due to a lack of knowledge. This intelligence, indeed, is situated in time and space. It is a subjective point of view on the world. All it sees are phenomena within this spatiotemporal structure. Yet, it seeks to see not the phenomena appearing within this structure, but the contours of the structure itself. Is not the problem of consciousness, understood in this way, as difficult as observing time without any event? Or observing the Universe in the absence of matter? Time only manifests through events. Space is only revealed through extended bodies. Without bodies, without events, nothing conceivable exists.

To offer you a metaphor, we could say that the unity of the body (a kind of spatial object) and the mind (a kind of temporal event) presents itself to us in the same way a cylindrical object would present itself to a two-dimensional being. For this being, the cylinder would appear either as a circle or as a rectangle, depending on whether it is projected from a longitudinal or transversal cut. In other words, our two-dimensional being would see two different objects, a circle and a rectangle, depending on the

[1] It is worth noting that Freud, a trained neurologist, sought to give his theory a precise physiological substrate. In his early essay, titled *Sketch of a Scientific Psychology*, he models the brain through principles inspired by thermodynamics. It must be acknowledged that his theory gradually detached from a purely naturalistic description.

adopted viewpoint. It would have no reason to believe, *a priori*, that these are two perspectives on the same three-dimensional object.

Just like this two-dimensional being, we see reality either through the prism of the mind or through the prism of matter, and we suspect that there is a link between these two perspectives. We therefore hypothesize (1) that mind and matter are two distinct realities and (2) that these realities meet at a point in the brain. We also believe that (3) the reality of the mind must be explained solely through that of matter, so that terms referring to the mind (thoughts, ideas, sensations, pain...) are ultimately nothing but physiological phenomena.

This view I have just described is that of reductionist materialism. In this chapter, I would like to suggest the very opposite; namely, (1) that mind and matter are one and the same reality, neither purely mental nor purely spatial, but (2) that there is no point of contact between mind and matter, and (3) that reality cannot be fully explained by either the mind or matter alone. I will argue that our understanding of the world requires us to approach both perspectives. Some problems receive an adequate answer from the angle of the mind (am I in pain? am I happy? am I conscious at this moment t?) and others demand the perspective of matter (is the Universe finite? how to produce energy?).

To make my philosophical position credible, I will first provide arguments showing that the mystery of the relationship between mind and body, and especially the brain and consciousness, presents an insoluble logical paradox. Every time we parallel a subjective experience of consciousness (perception, sensation, idea) with a physiological phenomenon (neural excitation), assuming that there is a causation between the appearance of an electrical signal and the birth of a thought, we commit a judgement error. This error is what I call a metaphysical chiasmus; it consists of mixing terms that have no relation to each other. More precisely, I will show that the paradox we encounter is a paradox of self-reference. I will also distinguish two types of metaphysical chiasmi: *internalist* chiasmi and *externalist* chiasmi.

However, I will not merely separate, with a stroke of the wand in the middle of the Red Sea, what pertains to the mind from what pertains to matter. I will show that, although we cannot think the relationship between consciousness and matter, it is impossible to conceive of consciousness without implicitly referring to matter; and it is equally impossible to conceive of matter without implicitly referring to consciousness. In other words, I will attempt to suggest that body and mind are like day and night. Without light, it is impossible to think of what a shadow is; but without shadow, there is no light. One cannot think the objectivity of matter without referring to the subjectivity of consciousness; and vice versa.

This collusion between consciousness and matter, this fundamental logical entanglement of concepts, will lead me to hypothesize that body and mind are merely two facets of the same substantial reality — the two faces of

an ageless goddess. They cannot be separated, as they form a single whole; but neither can their connexion be explained. The curse of human reason lies in abstracting terms from one another so thoroughly, in order to give them independent definitions, that it finds itself unable to think the unity from which these terms originated in the first place.

The mysterianism I have outlined is what I will call *perspectivist mysterianism*. The strategy of this chapter is plural. First, I will pose a few hypotheses about consciousness and show how these hypotheses lead to the famous paradox of consciousness. Then, I will show that this problem can be bypassed if we admit that body and mind form a coherent unity, but not one that is entirely explicable. In the following chapters, I will attempt to show that my philosophical vision is consistent with contemporary theories of the brain and consciousness.

12.2 Hypothesis 1: Living Consciousness and Dead Consciousness

From a purely intimate and introspective perspective, the only indubitable thing for a rationalist philosopher is to be, at a given moment, conscious of something rather than nothing. The tangible world that unfolds under the groping of their fingers, eyes, and tongue exists only in the consciousness they have of it. The very idea of being a subject perceiving the world, a body within matter, a thinking personality endowed with its history, memory, and immutable character, seems open to question. Only the present moment exists in its fullness.

From the point of view of scientific experience, on the other hand, it seems that the primary certainty is the existence of a material world, made up of hadrons, waves, and energy, within which it is difficult to explain how the phenomenon of consciousness as such emerges. When I look at the eternal sky, when I picture to myself, one by one, the infinity of worlds dotted across the vast expanse of the Milky Way, as so many stars adrift in the frozen void of the abyss, I cannot help but feel a curious pang of melancholy. The Universe seems terrifyingly more real than the consciousness I have of it. It is filled with sounds, colours, intoxicating fragrances, which are nothing more than quantum particles in stochastic motion. What I take for a true phenomenon of consciousness (the contemplation of the sky) is merely a bit of noise and fury (neurological reactions in the brain).

One might therefore think, as a first hypothesis, that consciousness can be defined in two distinct ways:

- In a *naturalistic* sense, which involves viewing it as a black box whose behaviours need to be simulated, reproduced, or predicted. The data gathered from experiments provides abundant information about the

specificities of consciousness as a function of the organism, but not about its nature as lived experience. In this regard, the approach can be described as external and objective. External, because it concerns the representable and quantifiable manifestations of consciousness. Objective, because the adopted viewpoint is a third-person perspective. Epistemically, it is assumed that the result of the experiment is not tainted by the subjectivity of the observer: the object is supposedly presented as it is, and not as it is seen. Some definitions or characteristics of consciousness found in the literature fall under this category, notably terms like "conscious creature[2]", "access consciousness", "monitoring consciousness[3]", and also in computational models of consciousness, whether based on attention[4] or information integration[5]. These models often rest on a reasoning impasse comparable to Hume's problem of induction in his HUME DAVID, *An Enquiry concerning Human Understanding*. With this understanding of consciousness, purely external and descriptive, science sometimes tends to reduce mental life to a set of biological processes concerning dead matter. For this reason, the naturalistic consciousness will sometimes be referred to, in a more caustic manner, as *dead consciousness*. This dead consciousness aims to provide a model of the brain.

- In a *phenomenological* sense, which involves viewing it as a principle of knowledge. This is the approach adopted in the Cartesian *cogito*, but also in spiritual experiences. By spiritual experiences, we mean all practices aimed at awakening consciousness to its inner world, or more precisely, to the idea that consciousness forms a totalising unity outside of which nothing can be conceived. This definition of spirituality is not far from that given by Foucault in *The Hermeneutics of the Subject*. Insofar as the knowing subject is also its own object of knowledge, the search for truth takes place within a process of self-transformation. All knowledge acquired about consciousness is knowledge acquired about what one is (or what one is not). It is thus an internal and subjective approach. Internal, because it concerns consciousness as it is in its intimacy. Subjective, because the viewpoint here is not that of God, but is a first-person perspective. Epistemically, it is assumed that the subject does not need to refer to anything other than their own experience to speak about their lived

[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

[4] TAYLOR, *Paying attention to consciousness*, Progress in Neurobiology, 2008.

[5] The Integrated Information Theory, defended by Giulio Tononi on the foundation of Gerald Edelman's idea, will be the core subject of the chapter 15.

> experience: one tends to forget that language, culture, and objective representations might, for example, play the role of filters or biases in understanding oneself. As a result, consciousness is often seen as a fact, where phenomena are given in an immediate and representational way. Some definitions or characteristics of consciousness found in the literature fall under this category, notably terms like "qualia", "phenomenal structure", and "subjectivity[6]", "self-consciousness" and "phenomenal consciousness" (Block, 1997), "introspective knowledge" and "homogeneity[7]", or "first-person perspective[8]". Insofar as this definition of consciousness takes into account the lived experience in its immediate appearance, but not the underlying structures of this experience, I will call this consciousness *living consciousness*. This consciousness aims to provide a model of the mind.

In light of the above, any mental content presents both a part of *observation* and a part of *introspection*. Introspection considers this mental content as a phenomenon of consciousness, disregarding the existence of a reality beyond the senses; observation, on the other hand, removes all that is subjective from a mental content and retains only the objects that are being targeted. If introspection retains from the thought of an apple that "I see an apple", observation, for its part, retains only this: "There is an apple". In the first case, knowledge evolves within living consciousness; in the second case, it targets a reality beyond the experience itself. To express this in Husserlian terms, introspection represents the phenomenal dimension of consciousness (i.e. the very act of feeling, thinking, experiencing), while observation represents the naturalistic dimension (the intentional object on which thought is focused).

The existence of an introspective life, a subjective lived experience from the first-person perspective, has often been taken as the touchstone of philosophy. This is evidenced notably by the Cartesian cogito, Hegel's sensible certainty, Kant's transcendental ego, or Leibniz's monads. However, it presents a difficulty whose implications need to be understood. When one starts from introspection, that is, from the "I" perspective, one faces the vivid phenomenon of consciousness, but the problem remains of determining the limits of this consciousness. Just as the fish, immersed in the Ocean, wonders what the water it swims in is and what its contours are, philosophy is tasked with determining the validity of the concept of "I". Throughout this thesis, I will hypothesize that the problem of the "I" is a problem of representation, which makes consciousness an open space, with no clearly

[6] Van Gulick R., *op. cit.*

[7] Lycan William G., 1999

[8] Nagel Thomas, *What Is It Like to Be a Bat?*, The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.

defined boundaries.

12.3 Hypothesis 2: The Paradox of Consciousness

The second hypothesis I would like to introduce regarding consciousness concerns the articulation between the naturalistic approach (the dead consciousness) and the phenomenological approach (the living consciousness). Before presenting this hypothesis, I would like to remind the reader of certain topological concepts that, I believe, will help in better grasping what follows.

In topology, a distinction is made between *open* sets and *closed* sets. An "open set," or more simply an "open," is a region of space bounded by a certain contour, but which excludes the points forming this contour. A concrete example of such a space can be obtained by considering a disk from which the boundary is removed. A closed set, on the other hand, is a region of space that includes the boundary.

Keeping these distinctions in mind, my second hypothesis will be to defend the idea that consciousness is an *open* space: although all reality exists within consciousness, the contours of consciousness itself cannot be represented within it. Indeed, how could one have a finite and bounded representation of the infinite and constantly renewing capacity to form representations?

This difficulty in consciousising consciousness can be illustrated with an analogy. Suppose an individual has grown up all their life in a room devoid of any reflective surfaces. They are capable of seeing the world around them: the walls, tables, chairs, etc. They can also contemplate their hands, their body, and — with a bit of luck — the tip of their nose reddened by an excess of alcohol, but they have never seen their own face, nor the colour of their eyes. How could such a person explain the phenomenon of vision?

Our hypothetical individual would certainly understand that vision, as a daily experience, must have an origin within themselves. They might even claim that this origin is somewhere on their face, but, although they can close their eyelids, touch their eyes with their fingers, they would not be able to see their eyes with their own eyes. In other words, vision would be, for them, a sensory obviousness, but not a visual obviousness. They would be aware of their own eye as something abstract, something they can touch with their hands, something they can feel within themselves, but that they cannot see.

What becomes of this analogy when we relate it not to the eye of vision, but to consciousness? Just as the eye does not fall under the sense of sight itself, it could be that consciousness, as an experience, is not graspable by the resources of our consciousness. Every time I represent myself as consciousness, every time I say "I am such," "I am the one who thinks,"

"I am cogito," "I am the subject of all experience," or "I am a brain in the world"; every time I engage in self-contemplation, I have before me an image, an emotion, a sensation of myself. This is an object of my thought, like many others, and does not differ in this from the representation of a table, a book, or a monkey. This image of myself is not my objective consciousness as such, but a mere subjective representation contained within my experience of consciousness.

This apparent paradox may be the fundamental reason why the conscious world seems populated with objects that are in me but are not *me*? The central idea of this second hypothesis is to think that consciousness, unlike ordinary objects, cannot be represented. It can either be *lived* subjectively as a "I" (I am consciousness), or *conceived* abstractly and objectively as a "He" (there is consciousness).

In the first case, consciousness is a sort of indefinable something about which nothing can be said. It manifests itself, but the moment it appears, it is already gone. In the second case, it is nothing but an abstract idea, one that cannot be concretely represented. This abstract consciousness is a bit like the proposition "Homo sapiens is the being who thinks he thinks." We understand that Homo sapiens thinks about the fact that he thinks, but we do not know exactly what he thinks about concretely. Thinking about thinking does not seem as visual as thinking about the appetising apple in the middle of my garden.

12.3.1 Consciousness is Unrepresentable

In accordance with our first hypothesis, the word "consciousness" would therefore refer to two distinct concepts. On the one hand, it would denote the succession of all possible experiences, and on the other hand, the unrepresentable idea of the "Self" that contains these experiences.

The first dimension is what I have called *phenomenological consciousness*. It always pertains to a specific content: a perception, an idea, an emotion... In this sense, we can say that phenomenological consciousness is a representational consciousness or, to use Husserl's terminology, it is an intentional consciousness. This consciousness is always consciousness of something. It is impossible to be conscious of nothing, just as it is impossible to represent nothing.

The second consciousness, *naturalistic consciousness*, exists beyond any determined mental content: it is the container within which conscious phenomena unfold. It is therefore a non-representational consciousness. This unrepresentability is neatly summed up by Sartre's aphorism: "one *lives* the interiority of consciousness, but one does not contemplate it[9]". One might

[9]The exact quote is somewhat less elegant, and I have taken the liberty of simplifying it for the sake of the argument: in Sartre, J-P., *La Transcendance de l'Ego : Esquisse d'une description phénoménologique*.

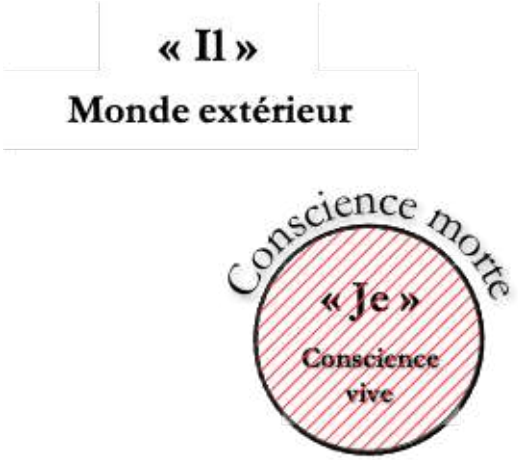


Figure 12.1: The circle (shaded area and contour) symbolises consciousness in its broadest sense. This is divided into two parts: living consciousness (shaded area) and dead consciousness (the circle's contour). The former consists of all the possible and imaginable conscious contents; the latter is the very fact of being conscious, i.e., of containing conscious contents. Two approaches are then possible: (a) The introspective approach starts from the "I" perspective and attempts to describe dead consciousness in phenomenological terms; (b) The scientific approach starts from the "He" perspective and attempts to describe living consciousness in naturalistic terms. Both the external world of the "He" and the internal world of the "I" are open spaces in the mathematical sense.

also say that dead consciousness is a hyperintentional consciousness because it stands beyond all possible and imaginable intentional aims.

To recap, the second hypothesis is to think that the difficult problem of consciousness, the one that aims to grasp the link between body and mind, is a problem of representation. More precisely, I would like to defend the idea that, at the root of the visible impossibility we experience in localising consciousness in the brain — and more generally, in material space — lies the impossibility, for a finite mind like ours, to form a representation of the infinite in act. That is to say, we cannot circumscribe in a finite thought the limitless fertility of consciousness, understood as the power to endlessly create new mental contents.

Note that when I say consciousness is "unrepresentational", I do not mean that it is "unrepresentable" *stricto sensu*. It goes without saying that dead consciousness, the one spoken of by science, can be represented in mathematical and neuroscientific language. It is a formally defined object, endowed with neural processes, mathematical functions, governed by verifiable hypotheses, and leading, in the best case, to predictions. However, these properties, representable in the mind of the scientist, are not *the experience of consciousness* that the scientist aims to describe. There is, therefore, symbolic representation of a consciousness that, by nature, lies beyond conscious representations. In a way, one could say that what science expresses about consciousness, it expresses in an analogical sense.

When Antonio Damasio or Gerald Edelman talk about the excitation of a group of neurons in the somatosensory cortex, they intend to designate the existence of a certain proprioception. They are of course aware that the language used to describe this sensation is not the sensation itself. Their scientific theories invoke representations that everyone has access to in their consciousness: the representation of neurons, synaptic activities, mathematical figures, etc. But none of these images are, as images, the dead consciousness as such: they are at most contents of living consciousness among others. No one would think of saying that a yellow object is yellow in and of itself, or that the drawing of a tree is the concept of the tree, would they? Science uses conscious representations that, in reality, point to a phenomenon of consciousness situated beyond conscious representations.

Of course, consciousness is not the only scientific object to maintain a relation of representation. When theoretical physics speaks of fundamental particles, the symbols e or p *represent* these particles — but everyone knows that the symbols themselves are not electrons or protons. There is, however, a fundamental methodological difference between the objects of physics and the singular object that is consciousness.

When I say that an electron has such an electric charge, such a mass, such a spin, I am merely speaking about a reality existing "outside" of consciousness. It is therefore taken for granted that the mental images I have of the electron are not electrons as such. According to the scientific ideal, I

should even bracket all my personal conceptions in order to understand my object of study as adequately as possible.

However, this neutrality does not hold in the case of consciousness. If I say that consciousness is "a phenomenon of the reentry of brain signals sent by the thalamo-cortical system[10]", I cannot help but compare these signals to the subjective contents of consciousness, since, by definition, consciousness is the sum of subjective contents.

This paradox of representation may be what explains why individuals, when they read a new scientific definition of consciousness, sometimes feel that this definition does not correspond to their lived experience? They will object that consciousness is not a mechanism, but that there is indeed a ghost in the machine: a "soul", a "mind" or a "vital principle". These common-sense resistances to the scientific demarcation must be taken seriously. They suggest that reducing representations of dead consciousness to living consciousness is a dubious operation, which I will call the *metaphysical chiasmus*. I will show that there are two types of such chiasmus.

12.3.2 Externalist and Internalist Chiasms

The study of dead consciousness is relevant when it comes to describing or predicting the behaviour of a conscious subject in a space of objectivity. The study of living consciousness, on the other hand, is pertinent when it comes to understanding or making sense of what we are as a perceptive whole. But as soon as one attempts to address questions regarding dead consciousness from the first-person perspective, or conversely, when one tries to penetrate living consciousness from the third-person perspective, one encounters methodological chiasms:

- **Externalist Chiasm:** The pitfalls of an approach in "He" become apparent when science, ceasing to focus on the external manifestations of consciousness, attempts to grasp its inner life. This approach is problematic, as there seems to be a petitio principii. Indeed, it is assumed that the measurements obtained within an experimental protocol refer to the subject's interiority, whereas they are in fact merely the result of a correspondence between what is observed on a strictly factual level, for example, through an electroencephalogram, and what the subject claims to feel at the moment these measurements are taken. Rather than establishing a causal relationship between an internal state of consciousness and an external manifestation, one is in reality only correlating two different kinds of external manifestations: physiological measurements and oral declarations. This difficulty does not stem from the imperfection of measurement tools, but from the fact that it

[10] This definition, as we will see later, is that of Gerald Edelman.

is impossible to "enter" someone else's mind and feel what they feel. It is a dead-end in reasoning comparable to the problem of induction as identified by Hume in his *An Enquiry concerning Human Understanding*, since the conclusion that is sought cannot be strictly proven rationally.

- **Internalist Chiasm:** The pitfalls of an approach in "I" become apparent instead when, from a first-person perspective, one attempts to express something about a reality outside of living consciousness. This gesture is impossible because, within this attitude, everything that is intelligible is part of the consciousness as a whole. The problem with living consciousness, to put it differently, is that it presents itself as a container that encompasses all phenomenality, yet its boundaries are impossible to discern. In our attempt to describe this consciousness, we find ourselves in a position comparable to that of a fish attempting to define what water is. Since water is its surrounding environment, it is impossible for the fish to step outside of this environment to describe it from the outside. Similarly, it is impossible to step outside of living consciousness and isolate it in order to categorise it as intelligible, unless one abandons the first-person perspective and, at the same time, deprives oneself of the experience of consciousness itself. It is therefore aporetic, within an "I" approach, to ask the question of what this "I" that thinks is: a substance? an object in the world? The internalist metaphysical chiasm is comparable, not to the problem of induction, but to the paradoxes of self-reference. It naively assumes that the subject, through introspection, has direct access to what they themselves are.

Having established these definitions, we will defend, throughout these pages, the idea that there is no such thing as pure subjective consciousness or pure objective consciousness, but that research in cognitive sciences is always situated within a crossed approach, and therefore methodologically problematic. To do so, we will first demonstrate, by studying the epistemic structures of scientific experiments, that each time a theory claims to say something about pure dead consciousness, it is not entirely in an exteriorising approach, but in fact conceals an externalist chiasm, such that the only possible conclusion, if it were to be rigorous, would be to say that living consciousness does not exist[11].

[11] Eliminativism is defended by Patricia Churchland (1986) and Paul Churchland (1981), as well as by proponents of illusionism. This even more radical idea posits that a materialism purged of its contradictions requires seeing what I have called living consciousness as an illusion. See, in particular, the work *Consciousness and Matter* by my former colleague François Kammerer.

Next, we will suggest, starting from the study of the Cartesian cogito, that each time living consciousness claims to speak of itself as an object in the world, it produces an internalist chiasm, such that the only possible conclusion, in its eyes, would be to say that everything is nothing but consciousness; or, which amounts to the same, that consciousness is nothing but an undifferentiated whole. In the first case, pure objective consciousness, forced to take refuge in its last strongholds, vanishes before the magnifying lens of the experimenter as if by implosion. In the second case, pure subjective consciousness, unable to see itself, fades away as if by dilation.

Since there is no such thing as pure subjective consciousness and pure objective consciousness, it will be necessary, in the second step, to ask where the idea of the exteriority of living consciousness or the interiority of dead consciousness comes from. It is one thing to say that these questions are devoid of meaning, but it is another to ask where they originate. If indeed these inquiries were truly contrary to rational thought, we would be justified in believing that they could never have been conceived. Yet, it turns out that they are indeed present in us, and so deeply nourish the scaffolding of language that it seems to be entirely supported by words referring to them: objective, subjective, interior, exterior, body, matter, mind, soul, etc.

This positions philosophy to undertake an inquiry into the genesis of these concepts. It does so with the aim of explaining, not what exteriority or interiority is, but what the very idea of exteriority and interiority is. In this respect, the approach can be described as phenomenological and transcendental. Phenomenological, because it evolves within the very phenomenon, but attempts to transcend the mere perceptive immediacy of lived experience through a language whose meaning refers to objectivities, or more precisely to the intentional aims of thought. Transcendental, because the question being addressed is what are the conditions that make possible the idea of the interiority of objectivity and the exteriority of subjectivity.

12.3.3 Links with Logical Paradoxes

I have described how the relationship between vivid and dormant consciousness is paradoxical, yet without formally defining the nature of this paradox. I have merely spoken by analogy, using the example of a man attempting to view his own eye. It is, however, legitimate to question whether the problem of representing consciousness is akin to known logical paradoxes, such as the Barber Paradox (section 5.2.1) or Cantor's Paradox (section 5.2.2).

Beyond clarifying my argument, this question arises from a more fundamental problem: how could it be possible to contemplate logical paradoxes (including the Barber Paradox) if consciousness were not paradoxical itself? Allow me to elaborate.

By definition, vivid consciousness encompasses all phenomena. This means that no thought, outside of consciousness, can be conceived. All

the objective knowledge we hold about the world, all our science, our genius, our art, the books accumulated over generations exist within human consciousness (albeit it's worth noting they may not exist solely within it. My point is that they will, in any case, exist within consciousness.)

For instance, our mathematical understanding of geometric space corresponds to a conscious experience that we can mentally reproduce, imagining bodies and subjecting them to translations and rotations. The same applies to our knowledge of numbers, bodies, and electrons.

If our consciousness of the world did not possess the structure of Euclidean space, it would be impossible for us to represent Euclidean space. Indeed, it is because our conscious experience does not have the structure of a four-dimensional space that we are incapable of concretely representing a four-dimensional hypercube or hypersphere. What I mean, in somewhat Kantian terms, is that the objects of our knowledge conform to the structures of our human mind.

The question that arises then is: from what experience of consciousness does our understanding of logical paradoxes emerge? If consciousness had nothing paradoxical or contradictory, how could we explain our ability to think about contradictory objects?

This line of thought allows for two possible answers. One could assert, as the mathematician Henri Poincaré once thought, that logical paradoxes are merely language games; abstract and unrepresentable constructs that have no existence beyond their linguistic formulation. This view was prevalent in the early twentieth century when the significance of the paradoxes of Russell or Cantor was not yet fully understood. However, it is now accepted that these paradoxes, far from being anecdotal, have major consequences on our conception of mathematics (refer to section 5.2.5 for further examples).

Alternatively, the second option is to admit that logical paradoxes correspond to representable realities of consciousness. This latter scenario aligns with dialetheism, which acknowledges the existence of paradoxical objects that are both true and false. Famous advocates of this position include Hegel, Łukasiewicz, and Graham Priest.

Even assuming dialetheism, it is nonetheless not entirely intuitive to explain how humans might represent a square circle. As soon as we visualise a square circle, an image comes to mind; but a simple inspection reveals that this image corresponds neither to the definition of a circle nor a square. In truth, it is impossible to form a concrete, sensible representation of a contradictory object; thus, it is easier to assert that such objects do not exist. Note that this reasoning applies to other types of paradoxical objects such as "the set of all sets" (Russell's paradox) or "the cardinality of all cardinals" (Cantor's paradox) or "the smallest integer not nameable in less than eighteen syllables[12]" (Berry's paradox).

[12]This paradox arises from the fact that the phrase "the smallest integer not nameable

It seems that paradoxical objects occupy an intermediate position between representable objects and those that do not exist: they are logical realities without explicit representation. We can conceive them abstractly, but we cannot sense, see, draw them in imagination—nor can we think of them coherently. Hence, it must be admitted that there exists within consciousness abstract non-representable objects. Such objects occupy the frontier of consciousness. They reside within consciousness because we can conceive them, yet at the same time, they do not, as we cannot represent them.

Now, what is the only object of consciousness that is conceivable without being representable? According to our second hypothesis, it can only be consciousness itself. The boundary of vivid consciousness, as mentioned earlier, is the objective idea of consciousness. It follows that the experience giving rise to contradictory objects—what makes our theory of paradoxes possible—is the experience of consciousness.

I can already hear the reader objecting that consciousness is not the only non-representable object. Figures in four-dimensional space, you might say, also possess this property of being conceivable but unrepresentable. The same applies to the abstract terms of our language. Who can concretely represent what justice is? An idea? What Gibbs free energy is? An infinite space? It seems our knowledge abounds with words without obvious representations, without needing to refer to consciousness.

These counterexamples, however, do not suffice to challenge my position. Neither justice, nor Gibbs free energy, nor even infinity are strictly paradoxical objects. They are certainly unrepresentable, but not contradictory objects. That is, we can think of them, assign them logical properties, describe relations they maintain with other concepts. Consciousness, on the other hand, is not logically thinkable. It is conceivable but not thinkable *stricto sensu*.

Consciousness in Relation to Russell's Paradox

In the following pages, I would like to more precisely characterise the paradox of consciousness, in order to understand the link existing between the concepts of vivid consciousness, dormant consciousness, and logical paradoxes. It is easy to conceptualise, albeit in a purely vague and intuitive manner, that the paradox of consciousness mirrors the structure of Russell's Paradox, also known as the Barber Paradox (see section 5.2.1). This paradox involves considering the set R of all sets that do not contain themselves; and questioning whether R contains itself. If it does not contain itself, it must contain itself. And if it contains itself, it does not contain itself.

The structure of Russell's Paradox seems akin to the paradox of consciousness. Vivid consciousness is indeed defined as the set of lived reality.

in less than eighteen syllables" itself consists of eighteen syllables.

Yet consciousness appears to be something lived. Therefore, consciousness must be within vivid consciousness. The paradox arises from the fact that if consciousness is within consciousness, it presents itself as a mere object of consciousness. Thus, it ceases to be a conscious subject. Consequently, consciousness cannot be within consciousness; yet it must. If consciousness is not within vivid consciousness, then it does not exist as lived reality, which seems to contradict the experience of being conscious.

It is important to note that this paradox is deeper than it may initially seem. Not only does it assert that consciousness is not something mental, as it cannot exist within vivid consciousness; but it also excludes the possibility of forming a scientific materialist representation of consciousness. If, indeed, we possessed a neurophysiological model of consciousness, we would have an explicit representation of consciousness. This representation would be a mental image of our consciousness. However, the paradox precisely excludes the possibility of creating an objective image of consciousness. At most, if we wish to circumvent the paradox, we can only assert that consciousness exists materially, or in other words, objectively, in a space beyond consciousness—and that we cannot form an explicit representation of this space. This leads, in more pragmatic terms, to one of the following alternatives:

1. Consciousness exists as a material reality, but there is no materialist theory of consciousness.

2. Consciousness exists as a material reality, but we cannot comprehend this theory.

These two alternatives, as a conscientious reader might note, coincide with the answers we considered for the dilemma of intuition in chapter 9. To recap, the dilemma of intuition was about how it is possible to give intuition an explicit form that nevertheless escapes a Turing machine's computational abilities. The problem, therefore, was to explicate, in an explicit form, why there is something inexplicable about intuition. This problem is fundamentally the same as that of consciousness, questioning how we can form a representable idea of something non-representable. The hard problem of consciousness and the dilemma of intuition are mere manifestations of a more fundamental logical paradox.

Consciousness in Relation to the Cheshire Cat's Paradox

Although there is an approximate analogy between the problem of consciousness and Russell's paradox, I would like to present a logical paradox that adheres more rigorously to the specific properties of consciousness. This paradox is of my own invention. I'd like to introduce it through a dialogue featuring two interlocutors. To pay tribute to Charles Dodgson, who greatly

contributed to the interactions between mathematics, philosophy, and literature, I shall take the Cheshire Cat as the questioner and Alice in Wonderland as the respondent. The paradox is arrived at through a dialectical exchange:

— Cat: Let's play a little riddle. I'm going to think of a random natural number. If you find it, you win.
— Alice: But how am I supposed to guess this number?
— Cat: Just try! I'll tell you if your number is larger or smaller than mine.
— Alice: Very well, in that case, I'll say... let's see... maybe 1000?
— Cat: That's a bit low...
— Alice: A million?
— Cat: Even higher...
— Alice: But am I at least getting warm?
— Cat: Not at all!

My "paradox" posits that the Cat will never stop responding "that's a bit low" no matter Alice's answer. And for good reason: there are infinitely many natural numbers. The proportion of natural numbers less than Alice's response, however large, is utterly negligible compared to the wholly infinite total of natural numbers.

Consequently, unless Alice is given (a) an infinite number of attempts (which implies both she and the Cat are immortal) or (b) witnesses a miracle, it is impossible for her to win the game. This "miracle" I refer to would be infinitely less probable than obtaining the complete works of Shakespeare by letting a monkey randomly type letters on a keyboard, as Borel imagined in the infinite monkey theorem of 1909. It would, in fact, be less probable than any event with even an infinitesimal probability of occurring. It would be a mathematically possible event but with a probability exactly equal to zero.

In a way, integers resemble an expanding universe. As space-time dilates, the atoms it contains drift further apart. Certainly, there are clusters of nebulas, but after a sufficiently long time, these will be mere isolated points in empty space. If the universe's expansion continues indefinitely, matter would eventually concentrate on its surface. And as this surface grows, the probability of finding a particle inside the Universe, rather than on its surface, would approach zero.

In the case that concerns us, the set $\mathbb{N}$ of integers is not expanding. What is expanding is the enumeration we can make starting from 0 and following a series of increasing numbers. However, we might represent, as a sphere, the set of numbers smaller than Alice's response, defining the radius of this sphere. As long as the sphere has a finite radius, the numbers within it are negligible. But as soon as the sphere has an infinite radius, it encompasses all numbers. Most of these numbers, moreover, seem to lie

on the sphere's surface; that is to say, most integers are more distant from the centre than any representable distance. It thus becomes paradoxical to comprehend where these numbers actually are. Are they within the sphere? Yes, they must be, being finite numbers. Yet they cannot surely be at any finite distance from the centre. Therefore, they are both inside the sphere and outside it.

I intend to argue, in the pages that follow, that the paradox of consciousness is exactly like the Cheshire Cat's number: objective consciousness is a self-extensive object; an object at the surface of a sphere whose radius stretches toward the horizon. To grasp this analogy, and circumvent some technical objections one might raise regarding probability theory, it is essential to first establish some rigorous definitions.

Formalisation of the Paradox

Définition 28. *Let $\mathbb{E} \subset \mathbb{N}$ and $n \in \mathbb{N}$. The density of E in $\mathbb{N}$ is denoted δ_E and is given by:*

$$\delta_E = \lim_{n \to +\infty} \frac{|\{e \in E \mid e < n\}|}{n}.$$

The density δ_E defines a measure of the measurable space $(\mathbb{N}, \mathbb{P}(\mathbb{N}))$. Intuitively, it captures the probability, if drawing a natural number at random from $\mathbb{N}$ (according to a uniform probability distribution), that this natural belongs to E.

Exemple 4. *Let $A = \{2, 4, 8, ..., 2n\}$ and $B = \{2, 4, 16, ..., 2^n\}$. The density of A in $\mathbb{N}$ is $\frac{1}{2}$. However, the density of B in $\mathbb{N}$ is zero. Both sets have the same cardinality, but in terms of density, one is infinitely superior to the other. Topologically, one might say that B is merely a point of A.*

Paradox 1. *Let m be any natural number. Draw at random a number X from $\mathbb{N}$ according to a uniform probability distribution $P(X < m) = \delta_{\{0,1,...,m-1\}}$. There is a probability equal to 1 that X is strictly greater than m.*

Proof. Firstly, we observe that the probability distribution is well-defined because $P(X \in \mathbb{N}) = 1$. However, for every $m \in \mathbb{N}$, the probability that $X \in \{0, 1, ..., m\}$ is given by:

$$\begin{aligned} \delta_{\{0,1,...,m\}} &= \lim_{n \to +\infty} \frac{|\{e \in \{0, 1, ..., m\} \mid e < n\}|}{n} \\ &= \lim_{n \to +\infty} \frac{m+1}{n} \quad \text{(as soon as n > m)} \\ &= 0. \end{aligned}$$

Therefore, there is a probability equal to 1 that $X \in \mathbb{N} \backslash \{0, 1, ..., m\}$, or in other words, that $X > m$. □

This "paradox" is counter-intuitive in that we might expect, for some sufficiently large natural number m, that it be probable, or at least not improbable, that $X < m$. But to think so is to forget that the number of natural numbers less than m is negligible compared to the number of natural numbers greater than m. This is true whether we count in terms of density, cardinality, or ordinality (see annexes 4 for an extensive definition of these terms).

Our finite minds, unfortunately, prevent us from imagining anything else, such that when asked to pick a natural number at random, we choose from a very limited set of possibilities. However, if we were to truly pick randomly, without any size constraints, this number would be very large. So large that it would be probabilistically impossible to propose a larger number.

Self-Extensive Objects

One might argue that the Cheshire Cat cannot choose a natural number "at random," as this expression is meaningless. In mathematics, indeed, one would need to define a probability function specifying the likelihood of choosing one number over another. Yet the probability function cannot be equi-probable for an infinite set. Its definition precludes it.

This objection is entirely valid, yet I believe it does not forbid us from conceiving, as a mere conceptual experiment, that the Cat might employ an equi-probable probability function. There is, in principle, nothing paradoxical about imagining such a function. We merely arrive at the conclusion that the number resulting from this random choice would have the peculiar property of being a natural number indeed—but it would surpass all others. Hence, it would not have a precise value.

Such a number would be what I term a *self-extensive* object. Each time an attempt is made to ascertain its precise value, by posing a question (are you within a given interval?), we would certainly gain additional information about this number. But this information would never allow us to place the number within an interval of finite size.

For instance, one might have an answer to the following questions: are you an even number? are you a multiple of three? do you admit an integer square root? These questions would receive clear and precise answers. Yet there would always remain an infinity of possibilities. In other words, no combination of questions exists that would position this self-extensive number within a finite interval.

In this respect, self-extensive numbers behave similarly to objective consciousness. We may ask questions about them and receive answers but without ever dispelling all unknowns remaining in their regard. For example, we might say it is necessary for the Cheshire Cat's number to be greater than a given integer n; or that it be a multiple of three. But we cannot find conditions both necessary and sufficient. It follows that these

self-extensive objects inherently conform to the definition of Jesse Prinz's mysterianism, which I discussed in section 9.4.4, and which he summarised with considerable clarity[13]:

> *One can know that drinking two litres of alcohol suffices to cause drunkenness, and that at least one teaspoon of alcohol is necessary, without knowing the amount both necessary and sufficient. In the case of consciousness, the properties we know to be sufficient are properties computers lack. The properties we know to be necessary may be possessed by computers, but we do not know if they suffice.*

This detour through the Cheshire Cat's paradox enables me to summarise hypothesis 2 in the following form:

Hypothèse 5. *Objective consciousness, along with all paradoxical objects, are self-extensive objects. Each time an attempt is made to characterise them, they extend beyond any representation we can provide. Certainly, it is possible to gather information about them, but this information will be negligible. From a geometric standpoint, they occupy the boundary of an ever-expanding sphere.*

12.4 Hypothesis 3: Dual-Aspect Monism

The duality of consciousness, which can at times be considered as living consciousness and at other times as dead consciousness, leads me to formulate a third hypothesis: namely, that there exists a tangible reality at the origin of dichotomous concepts such as "thought" and "matter," or "objectivity" and "subjectivity," but that these concepts fail to capture this reality in a way that is both complete and coherent. The reasons that drive me to think in this direction are manifold.

First of all, it is observed that each of the two registers, namely "I" and "It," can only provide partial insight into the question of consciousness, never offering the final word. They inform us sometimes (i) about the intimate life of the subject and the connexions between its contents of consciousness, and sometimes (ii) about the elements of matter and their causal relationships. They, however, speak much more difficultly about the relationship between the two aspects. It is suspected that there must indeed be a link, or a common structure, that explains the correlations observed between certain brain activities and self-reported mental life, but as soon as one attempts to explain a phenomenon starting from one of these registers, using concepts

[13] translated by me from the article *Level-headed mysterianism and artificial experience* by Jesse Prinz.

derived from the other register,

What is gained in completeness comes at the cost of a certain incoherence in the terms, made evident by the presence of symptomatic linguistic expressions. This is exemplified, for instance, by the mereological errors found in the writings of some neuroscientists, when it is said that the brain "decides" (Damasio[14]), that it "believes," "interprets," "guesses" (Crick[15]), that it "knows," "reasoning," "constructs hypotheses" (Blackmore[16]), that it "categorises," "distinguishes," or "manipulates rules" (Edelman[17]), or when it is asserted that neurons themselves are "intelligent" or "estimate probabilities" (Blackmore[18]). While useful as analogies, these expressions often betray the presence of an externalist metaphysical chiasmus, as intentional states are directly attributed to the external, or physiological, manifestations of the brain. This kind of error is also referred to, in the English-speaking literature, as the "mereological fallacy[19]".

Contrary to what some authors claim, we do not think that this chiasmus occurs *solely* because of language abuse—and that it would be sufficient in some way to purify neuroscientific theories of certain terms to resolve the problem *ipso facto*[20]. We wish to offer, against this view, two justifications.

The first is that if the analogy established between mental facts and physiological data were purely linguistic, the propositions about consciousness would lose much of their meaning, at least insofar as they would no longer truly illuminate the causes of the existence of this mental life we possess. This move, which essentially evacuates the error of reasoning along with the problem the reasoning sought to address, leaves us with a sense of dissatisfaction, as it puts on hold the neuroscientist's object of inquiry, their sincere and unspoken aim: to account for the obscure link between body and mind.

The second reason why it is difficult to explain these fraudulent expressions in purely linguistic terms is that one risks awkwardly relying on a form of positivism, or worse, assuming that there exists a fixed boundary between the world of words and the world of things. The very existence of a supposed

[14] Damasio A., *Descartes' Error*, 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, London 1997, p.91.

[17] Edelman G., *Bright Air, Brilliant Fire*, Penguin, Harmondsworth 1997, p.109

[18] Blackmore C., *Ibid.*

[19] Daniel Bennett M., Dennet D., Hacker P., Searle J., *Neuroscience and Philosophy: Brain, Mind, and Language*, Columbia University Press, 2003.

[20] Simon Ullman argues that the psychological properties attributed to the brain belong only to a certain technical vocabulary (cf. *Tacit Assumptions in the Computational Study of Vision*, p. 314.), while Richard Gregory and Ernst Gombrich argue that they are analogies (*The Confounded Eye*, p. 51.), and Blackmore speaks of metaphors (*Understanding Images in the Brain*, pp. 257–83.)

"language abuse" in fact means that a certain word is erroneously transposed to a domain of objects to which it should not refer. It follows that, on the one hand, the domains of mental objects and physical objects must be defined prior to language; and on the other hand, the meaning of propositions must be brought back, through a carefully constructed correspondence game, to these signifying objects.

This implies first defining the domain of possible objects, the semantics, and then inferring the domain of valid linguistic expressions, i.e., a syntax. Now, not only is the distinction between synthetic and syntactic propositions not easy to establish (this was the subject of the first part of this thesis), but it also seems that it is through the play of language that we seek to produce them; so that words, to some extent, condition the semantics we seek to attach to them. Given these difficulties, is it not reasonable to take the linguistic expressions that common sense and neuroscientific practice have provided us at face value and, rather than seeking to disqualify them, simply ask the following question: by what set of circumstances can these seemingly abusive expressions be endowed with meaning?

Thus, the task is not to dismiss, like an inquisitor, the propositions that are meaningless, but rather to seek the conditions for their intelligibility. Such an approach would have the consequence of returning to the concepts of mind and matter, of objectivity and subjectivity, and of questioning whether it is possible to talk about dead consciousness without simultaneously talking about living consciousness. Or, in other words, whether it is conceptually possible to isolate one from the other, or whether, on the contrary, these concepts are so intertwined that any expression dealing with one contains, in its very conditions of intelligibility, a reference to the other?

To provide a tangible illustration of this viewpoint, it may be helpful to keep in mind the simple example of the concepts of "inside" and "outside." When, in the most obvious topological use, I formulate a sentence such as "A is outside B," the very meaning of the sentence assumes that it is possible to understand the conditions under which A would not be outside B, that is, under which A would be inside B. Thus, I can only define what is "outside" in reference to what is "inside," so that it is impossible to have a discourse that completely separates the two concepts from each other.

Of course, knowledge of what is outside B does not imply a complete, or even partial, knowledge of what is inside it, but only of what it would mean for A to be inside B. This nuance is important because it amounts to saying that it is not the objects outside B that are intertwined with the objects inside B, but only the concepts of inside or outside themselves, that is to say, in other words, the conditions of possibility for an object to be inside or outside B.

In exactly the same way, complete knowledge of everything that can be said to be true, or could be said to be true, from an "It" perspective does not imply complete, or even partial, knowledge of everything that could be said

to be true from an "I" perspective; and it is not even necessary to assume a causal connexion between the objects of these two perspectives. On the other hand, understanding that a thing is true from the first perspective would consist in understanding what might render this thing false.

Now, if the "I" perspective and the "It" perspective are to be considered as concepts of inside and outside, as we have discussed above—in a way that still needs to be elucidated—it means that understanding what makes a proposition true in the "It" perspective (the outside) requires referencing the conditions of truth in the "I" perspective (the inside). To put it differently, our hypothesis would be to assume that the perspectives of "It" and "I" are conceptually intertwined in such a way that one can only be made intelligible in light of the other.

While it is easy to see how dual concepts such as "inside" and "outside," "cold" and "hot," or "big" and "small," can be conceptually intertwined, it seems more complex to perceive this in the case of dead consciousness and living consciousness. The former, indeed, seem nothing more than a formulation, sometimes positive, sometimes negative, of the same phenomenon, exactly like when one talks about a "half-empty glass" or a "half-full glass."[21]

However, in the case of dead consciousness and living consciousness, the intertwining cannot be thought of in terms of a concept and its opposite, but it remains plausible (or conceivable) if one can show that the "It" perspective, at a more fundamental level, invokes the "I" perspective. How exactly? That is what we must now ask, through four analogies.

12.4.1 First Analogy: Space and Extended Bodies

In an attempt to understand the entanglement between two concepts, the history of philosophy has provided us with a number of inspiring examples. One might then ask, to begin with, whether the relationship between the perspective in "I" and in "He" is not akin to the relationship between space and the bodies it comprises?

At first glance, space and the bodies that inhabit it appear to be two distinct concepts; yet, it seems just as difficult to imagine an extended body outside of any space, as it is to grasp a space without bodies. According to Kant, in fact, every external phenomenon summons, at a fundamental level, a certain pure intuition of space, which is necessary and logically prior to it.

In contrast to this initial result, a central idea of contemporary physics is to argue that space, far from being an ethereal container that lives in-

[21] A remark by Harald Atmanspacher on the concept of complementarity could be inserted here: "Let us note, however, that in the context of Boolean logic, the complementarity of a proposition differs from its negation. Relevant examples include the complement of belief, namely doubt rather than non-belief (James), or the complement of goodness, namely fairness rather than malignity (Bohr)." (Atmanspacher Harald, *Dual-Aspect Monism According to Pauli and Jung*, 2014-1, N. 3, pp. 105-135).

differently to its content, would ultimately be nothing but a product of matter[22]. This reversal is interesting because it suggests that, from an *a priori* perspective, space can be considered a necessary condition for the sensible knowledge of bodies; whereas from an empirical perspective, it is the bodies that, on the contrary, define space.

If this is indeed the case, we would have here a conceptual entanglement much closer to the one we are searching for, and the question would then become whether what we have called the dead consciousness, which is nothing other than consciousness considered from an external perspective, can be conceived independently of the possibility of phenomenological consciousness.

It is worth exploring this analogy and seeing whether it can teach us something about the very nature of concepts. This requires, at first, demonstrating how the perspective in "I" can be conceived as a necessary condition for the perspective in "He", and then showing the reverse relationship.

Deductive Step 1: The Concept of Living Consciousness as the Condition for the Possibility of the Concept of Dead Consciousness

Let us approach this idea carefully and reason step by step. Consider the case of a scientist who, taking the problem of consciousness from a naturalistic standpoint, would carry out the following experiment: a certain subject A states what their state of consciousness E is (their fears, thoughts, emotions, or representations...) when faced with a certain stimulus S. At the same time, the experimenter records the measurements M registered by an electroencephalogram.

The aim is to identify cause-and-effect relationships, or at least significant correlations, between the measurements M and the states E, while ensuring that these states remain stable in the repeated presence of the stimulus S. As an example, the stimulus S might consist of a prompt (think about your favourite sport), and the measurements M might be the activity observed via an electroencephalogram in the motor and somatosensory cortex. The states E, in turn, consist of asking the patient what thoughts or

[22]This proposition can be understood in two different ways, depending on whether one adopts the perspective of the theory of relativity or that of quantum mechanics. One might perhaps say, with Michel Paty, that "The space of the relativistic theory of gravitation (general relativity), structured by the fields generated by the masses of bodies, was no longer thinkable in the absence of physical bodies. One could say, in a sense, that empty space is constituted by the properties of bodies outside the place of their location, that it is the material place of the influence of bodies without the bodies themselves. As for quantum physics, it makes of space a material substrate in a latent state that can manifest at each of its points in the form of tangible physical phenomena (particles and fields) as soon as a certain degree of excitation is applied at that point (vacuum polarization and pair creation)" (Michel Paty, *Le vide matériel, ou la matière crée l'espace*, Simon Diner and Edgard Gunzig, Univers du tout et du rien, Éditions de l'Université Libre de Bruxelles/Complexe, p. 22-44, 1998.)

emotions accompanied the prompt.

The question we address is whether the experimental setup described above could achieve its objectives if the experimenter did not have the background, analogical idea, according to which the states E reported by subject A are similar to the states that they themselves would *be likely* to experience. I emphasise the term "likely" here because the question is not whether the experimenter actually experiences the same thing as patient A, nor whether they would report the same kind of sensations E in response to the stimulus S. The question is whether the experiment would still make sense in the absence of the joint possibility of a first-person perspective.

To attempt an answer to this query, and at the same time to clarify its contours, let us suppose that the same experimental setup we have described is carried out by a subject incapable of any intentionality — this kind of creature is commonly referred to as a philosophical zombie, or alternatively, by a machine equipped with sophisticated sensors. In this case, it would be hard to imagine that the zombie or machine could acquire the same understanding of the experience they are undergoing as an intentional subject would. And for good reason: the signals they produce and observe might be described by formal rules or predicted by models, but to them, these signals would have no more meaning than those emitted by an electron.

Such a zombie would, in fact, be reduced to a state akin to that described in Searle's Chinese Room. Everything, for it, would merely be the manipulation of symbols. These symbols might indeed refer to objects in the world, but not to mental states it could experience. The very idea that these signals might be the manifestation of an inner life could emerge only in an entirely abstract and intellectual way in its mind — if it even has a mind in the proper sense. It would appear much like our idea, as human beings, of what the life of a plant should be. One might imagine that the plant feels something, but the term "feel" would have meaning only in reference to our own sensory experience, which is undoubtedly very far removed from that of plants.

Our point here is not simply to say that an experiment, in as much as it always involves, at some stage, an interpretation, cannot be fully dissociated from the thoughts and mindset of the experimenter. A thesis of this kind is generally seen as a platitude and would not significantly advance our argument. What we argue more fundamentally is that the experiments we have described here, and which concern consciousness, involve a form of interpretation distinct from other experiments conducted in the natural sciences.

To clarify the point and simultaneously highlight the specificities of the empirical study of consciousness, it is useful to consider, as an archetype, an experiment in classical physics. When Galileo uses inclined planes to study the fall of bodies, he seeks nothing other than to establish relationships between different measurements, namely the time that elapses between the

moment when the object, starting from a given initial position, reaches the end of its trajectory, as a function of the incline of the slope. This experiment does not require reference to the subjective mental states of a subject. While it is true that the meaning of the experiment depends, to some extent, on the representations the scientist has of motion, time, or space, the theoretical terms, in themselves, do not refer to the scientist's mental states.

In the case of the neuroscientific setup we have described, however, there is the presence of a *theory of mind*[23]. The experimenter is not merely concerned with the patient's consciousness as a black box, into which certain information is input (the stimulus S) and certain other information emerges as output (the measurements M and self-reported states E). In reality, they are trying to infer something about the subject's "inner" experience. To understand the meaning of their conclusions, it is thus essential to establish an analogy between what I, as a subject, am potentially capable of feeling in the first person, and what I think the patient in the experiment claims to be feeling. This difference is precisely what differentiates today's neurosciences from the behaviourist paradigm that reigned during Alan Turing's time and the invention of the test bearing his name.

In conclusion to this reasoning, we can say that the "He" perspective invoked in a neuroscientific model of consciousness seems to acquire all its meaning only because of the prior possibility of the "I" perspective. Behind the objective representations in the third person, there always lies the necessary idea of a possible mental life.

Deductive Step 2: The Concept of Dead Consciousness is a Condition of Possibility for the Concept of Living Consciousness

Let us now consider the relationship in the opposite direction. In addition to what we have just said, it may well be that the first-person perspective cannot be made intelligible without the coexistence of a third-person perspective. Is it indeed possible to resort to introversion without indirectly referring to concepts that exceed introspection? Is pure introversion even conceivable? Or, in other words, can we have direct access to our own subjectivity?

Strangely enough, this critique seems easier to formulate than the one concerning the previous paragraph. Perhaps this is because of the recent history of psychology, and the unfortunate fate reserved for Wundt's methodology[24]? Perhaps, following Kant's line of thought, one could say that while

[23] In the sense of David Premack and Guy Woodruff when they write: "An individual has a theory of mind if they ascribe a mental state to themselves or others" (see the introduction to their article *Does the chimpanzee have a theory of mind?*, Behavioral and Brain Sciences, Vol. 1, N. 4, December 1978, pp. 515-526).

[24] As Arnaud Dewalq rightly states: "Wundt's idea is as follows: by virtue of the causal connexions between the psychological and the physiological, the study of physiological processes is regularly referred to data of internal observation, thus to psychological data."

all psychology begins with introspection, chronologically speaking, it does not logically derive from it?

Another way to approach the problem would be to take seriously the possibility of knowing oneself as a thinking substance and to see what dead ends this might lead us to. There is a certain paradox in the idea of the Cartesian cogito, pointed out by many commentators[25], which consists in the fact that if one thinks of phenomenal experience as the property of a certain subject, one immediately represents this subject as an object in the world. The thinking substance is thus conceived in a spatiality, which seems contrary to its very essence.

Or to put it another way, every time I attempt to represent what I am as pure intentionality, I incidentally invoke the idea of myself as something substantial. This does not mean that I am indeed an object in space, nor that space exists outside the thought we have of it, because such questions belong to a different order; but it does imply that the concept of an object, or of being oneself an object, is a condition of possibility for the concept of thinking substance. Could one indeed give meaning to the property of being something thinking if that something could not be conceived as an object having the property of thinking?

In the cogito, there is the presence of a kind of paradox: on the one hand, "I think" is the totality of thought, but on the other hand, it thinks of itself as an object of thought. It is as though it is both the container and the content, the one who thinks and that which is thought. This bifocal character points to the difficulty identified by Kant of distinguishing formal intuition as a subjective faculty (metaphysical space) from the object known through the exercise of this faculty (geometric space). It is as if the "I think" must produce proof that it is something other than the very content of thought. What exactly? A subject, no doubt, but in what sense? In the sense of being this kind of meta-property which consists not in conceiving this or that particular thing, but in conceiving things in general.

Now, for the cogito to think of itself as a property, it must refer to itself as an objectivity. This first implies differentiating itself from the world around it and from the ideas that pass through it, then recognizing perceptions as

(cf. *Deux théories de l'analyse psychique : Wundt et Brentano*, in *Vers une philosophie scientifique : Le programme de Brentano* [online]. Paris: Demopolis, 2014)

[25]Kant, first and foremost, criticises Descartes' conclusions, and those of rational psychology in general, in his chapter on the paralogisms of reason. He writes about the idea of the cogito that it is an idea of pure reason without real counterpart: "As excellent as the consciousness we have of our representation of these things may be, it is still far from the case that, if the representation exists, the object corresponding to it also exists..." (*Critique of Pure Reason*, p. 382). From this critique, it follows that, for Kant, there is no intellectual intuition. Only objects falling under the regime of pure sensibility possess an *a priori* form: space and time. For further development, refer to Auchatraire Anne, *Le cogito cartésien et la Critique de la raison pure*, in *Figures du Transcendantale*, ed. Crampe-Casnabet, Michèle, Lyon: ENS Éditions, 1996, pp. 41-63.

having a kind of permanence, and finally gradually acquiring a concept of object. But as it is difficult for it to distance itself from itself, and to look at itself from the outside, no doubt the subject must resort to all kinds of mediations? To representations, in a sense foreign to itself, on which it can base its judgement? Perhaps this is why authors like Husserl or Mead, who take seriously the hypothesis of a constitution of the self, insist on the importance of intersubjectivity.

This phenomenon, moreover, is observable even in quite simple situations from ordinary life, such as when we try to remember what we were thinking the day before, and suddenly we see ourselves there, in the middle of the room, as though through the eyes of a third person, leaning on our desk or on some wobbly chair, thinking about that thing we want to remember. The very content of our thought (e.g., the thought of a cat) often seems to be associated with the memory of the place where we were (I was thinking about this cat when I was on my desk chair), and with the position we occupied (I was sitting), at the moment of thinking it.

Of course, this is just a sensible metaphor, not an apodictic reasoning. But it suggests that when we try to access our own interiority, to that lively world of dreams, perceptions, and memories, we also often refer, *at the same time*, to the *possible* gaze of another individual upon us. That is to say, we see ourselves as a third party might see us.

Upon further consideration, it seems difficult to think of oneself as the subject of the thoughts, sensations, and perceptions that pass through us, that is, as the origin of all conscious content, without simultaneously seeing oneself in an objective and external way. Or, to reprise our cherished metaphor, perhaps one could say that thinking intentionality outside of an objective space is as difficult as thinking of an extended body outside a *possible* extended space.

The Two Concepts Are Therefore Conceptually Inseparable

This brief detour in our reasoning, which has led us from the paradox of the cogito to philosophical zombies, has only one purpose: since one cannot easily think of the "I" without invoking an objective representation; and since, conversely, one cannot describe consciousness in a theory of the "It" without invoking, within this very theory, an analogy to our own mental life, it seems that the concepts of objectivity and subjectivity are not entirely separable conceptually.

Rather than attempting to treat them as two distinct domains, or reducing one to the other, I suggest, in accordance with our hypothesis, that we consider them as the complementary, albeit paradoxical, faces of an underlying pre-conceptual reality. It is as if there is, in the end, only one kind of consciousness, a living-dead consciousness; but one that can only be thought of in two modalities, either as subject or as object. This conscious-

ness would not be material or ideal, strictly speaking, but would present itself more as an intuitive, pre-conceptual structure that our linguistic and scientific formalism could only apprehend at the cost of a certain methodological reduction.

Let us note that the reasoning we have just established in no way proves the presence of an external reality to our senses, as it can be reproduced coherently within the very transcendental phenomenological regression of Husserl or Descartes. This is, in fact, the result reached by Husserl when he concludes, in the Cartesian Meditations, that consciousness is always consciousness *of something*.

Moreover, it would be contradictory, given the initial premises of our approach, to assume or refute the existence of an "external world." To try to show that there is something beyond our senses, from an ontological point of view, is to risk falling into a form of dualism, for to have something outside of thought, there must first be something other than thought.

For this reason, we prefer to say that, from a certain point of view, everything is in thought, but that within this thought, there is an idea that refers to something from which it seeks to separate itself. We say "from a certain point of view" and not "absolutely," because there is yet another perspective according to which, on the contrary, everything is matter and phenomenal consciousness does not exist. This point of view, that of eliminativism, is not, strictly speaking, contested by our argument, but it is put into perspective: does the idea that thought emerges from matter not rely on a prior experience of consciousness?

This perspective we adopt amounts to asking the one who explains to us that phenomenal contents are illusions, where this conscious idea he nurtures about phenomenal contents comes from: is this idea not already a phenomenal content? Our thesis does not aim to invalidate the eliminativist viewpoint in terms of its conditions of *truth*, but rather to emphasize the presence, in the very conditions of *intelligibility* of the proposition that "phenomenal experiences are illusory," of a necessary reference to something that is phenomenal.

Let us also note, before we conclude, the presence of a final parallel between the concept of space and the perspective of the "I". From the *a priori* point of view, which was Kant's, it is somehow easier to represent space without extended bodies than bodies without space, since space is what makes the body conceivable. On the other hand, from an empirical point of view, the concept of the body seems primary in relation to that of space. Mathematicians such as Lobatchevsky or Poincaré even went so far as to say that our concept of space is determined by the experiences we have, and therefore primarily depends on bodies[26]. Now, concerning the

[26]Think, for example, of Lobatchevsky's empirical conception of space, where space is ultimately a mental construct obtained by the imagined, infinite juxtaposition of several

body-mind problem, it seems that in idealist approaches, the existence of the cogito is more easily accepted than that of external objects, whereas from a strictly empirical point of view, one would more easily accept the idea of external objects than the existence of one's own consciousness.

The relationship between the perspective of the "I" and the perspective of the "It" is, however, not of the same order as in Cartesian dualism, where the *res cogitans* and the *res extensa* are entirely different in nature and stand impervious to one another. Instead, their relationship is like that of Ying and Yang. The "I" approach privileges the phenomenological, while relying on naturalism. In the same way, the "It" approach starts from naturalism but takes shape thanks to phenomenal consciousness. It is this inherent hybridisation that represents both the best objection to physicalist monism and the best way to escape a dualism with no way out.

12.4.2 Second Analogy: The Incompleteness Theorems

At this stage, the reader might feel our third hypothesis has become difficult to uphold. How can one conceive of a monism where the two modalities are mutually exclusive yet necessary for understanding the whole? As strange as it may seem, this idea originates from an intuition concerning the nature of logic itself, particularly the incompleteness results of Gödel, Turing, and Tarski.

Recent results in predicate logic suggest that in a theory satisfying Gödel's theorems, the likelihood that a proposition with n symbols is undecidable asymptotically approaches one as n tends towards infinity[27]. From this result, knowing that there is no bound to the size of a well-founded proposition, it has been shown that in any formal system rich enough to axiomatise arithmetic, there exist infinitely more unprovable propositions than provable ones. In fact, the probability that a true formula is not derivable in propositional logic approaches 1 as the size of the formula, in number of symbols, tends towards infinity. Moreover, the set of demonstrable propositions, compared to the set of undemonstrable ones, is topologically akin to a cloud of points on a plane.

Of course, a proposition undemonstrable within a formal system S may well be provable in another system S'. The undecidability referred to here is not inherent but strictly concerns the axiomatic framework considered. This is why it is pragmatically effective to embrace a pluralism of theories.

---

bodies next to each other: "When they touch, two bodies A and B form a single geometric body C, where the components A and B remain distinct and are not lost within the whole of C. Conversely, an arbitrary *section* S divides a body C into two parts A and B... In this sense, all bodies in nature can be imagined as parts of a whole we call *space*." (Lobatchevsky, *New Principles of Geometry with a Complete Theory of Parallels*, Kazan, 1835-1838.).

[27] Cf. Calude, Jürgenstein, *Is complexity a source of incompleteness?*, Appl. Math. Comput. 66, pp. 63-76, 1994.

Since a single theory is likely to encounter, in principle, far more resistant propositions than decidable ones, it seems reasonable to believe (1) that no single formal system will account for the truth of all worthwhile propositions; (2) that the same problem may be undecidable in one formal system yet not in another.

The contemporary evolution of mathematics also seems to highlight the limitations inherent in founding the discipline on a single axiomatic theory. Although the Zermelo-Fraenkel theory has gained popularity, choosing axioms in such a way as to prove all interesting results, and only those, remains challenging.

Similarly, theoretical physics faces difficulties in unifying relativity theory and quantum physics. It may be that physics does not adhere to the incompleteness theorems of first-order logic, or that the issue of theoretical unification does not stem from the existence of undecidable propositions. Nonetheless, the coexistence of multiple methodological and disciplinary insights on the same problem has become common practice, if not desirable, in many fields of inquiry—as evidenced by cognitive sciences, which combine philosophy, linguistics, psychology, neuroscience, anthropology, and computer science.

Thus, there is no more contradiction in conceiving the existence of a monism with dual aspects than there is in conceiving the entire edifice of mathematics. Ultimately, it would be nothing more than a structure composed of a wide variety of more or less formalised intuitive contents, which could be explained through two competing axiomatic systems. One would start from the axiom of the *cogito*, and the other from the axiom that everything is matter. Nonetheless, they would indeed speak of the same objects.

In fact, the results of these two systems would coincide on a vast set of questions, indeed the majority of those constituting current human knowledge, but would also stumble upon an infinite number of undecidable problems. These would mainly concern meta-properties, such as whether the "He" perspective can address all the issues formulable within this perspective—a problem essentially related to the reductionism in natural sciences.

12.4.3 Third Analogy: The Principle of Complementarity

This specific relationship we attempt to describe between the "I" perspective and the "He" perspective may receive further illumination, this time inspired by quantum physics. In a 1952 article, physicist Wolfgang Pauli, known for establishing the exclusion principle, compares the mind-body problem to the wave-particle problem[28]. Rather than considering the mind as resulting from matter, which involves cross-causality, or assuming, like Leibniz, a

[28] Jung Carl Gustav, Pauli Wolfgang, *Naturerklärung und Psyche*, Zurich Rascher, 1952. For a more thorough philosophical discussion of their article, refer to *Atmanspacher Harald, Le monisme à double aspect selon Pauli et Jung*, 2014, N. 3, pp. 105-135

harmony between the order of thought and extension, he prefers to resort to an extrapolation of Bohr's principle of complementarity[29]:

> The general problem of the relationship between psyche and matter, between the inner and the outer, can hardly be considered resolved by the term "psychophysical parallelism" advanced in the last century. Nonetheless, perhaps, modern science has brought us closer to a more satisfying conception of this relationship, as it has developed the notion of complementarity within physics. It would be most satisfying for matter and psyche to be conceived as complementary aspects of the same reality.

The complementarity referred to here is not related to the non-commutativity of quantum operators, which lead to the violation of Bell's inequalities or the superposition states of the wave function. It rather concerns the wave-particle complementarity, where the same phenomenon admits two mutually exclusive descriptions, yet both are necessary. These two descriptions are required for a comprehensive understanding of the phenomenon, but they are not simultaneously accessible. According to Velmans, it is in this sense that the experiential life of a particular person and the brain activities of that person, observed by a third party, can be said to be complementary[30].

In extensive correspondence with Carl Jung, Pauli attempts to systematise this approach based on the idea that measurement in quantum physics achieves a precise yet imperfect partitioning of the phenomenon considered in its entirety—thereby depriving the observed objects of their highest connectivity. His fundamental intuition can be formulated, in algebraic quantum theory, using ontic and epistemic states.

Ontic states characterise the system as it exists in its totality, but they are not observable as such. Epistemic states, however, relate to a particular empirical context and can thus be subject to observation through an operational protocol. The problem arises from the empirical context breaking the high connectivity of ontic states. Objects emerge from measurement that no longer necessarily capture all facets of the studied object. This suggests there is a reality whose nature is neither purely mental nor physical, yet can be observed through these two viewpoints.

Pauli thus proposes establishing an analogy with how reality presents itself, or rather "is measured," in quantum physics, and the role of the unconscious in Jungian theories and, more generally, in psychoanalysis[31]:

> "[...] the epistemological situation regarding the concepts of 'consciousness' and 'unconscious' seems to present an analogy

[29] *Ibidem*, p. 164

[30] Velmans Max, *Understanding Consciousness*, second edition, Routledge/Psychology Press, 2009.

[31] Carl Jung, Naturerklärung und Psyche, Zurich Rascher, excerpt from *Atmanspacher Harald*, *Ibidem*.

> very advanced with the situation of "complementarity" in physics outlined hereafter. On the one hand, surely, the unconscious can only be detected indirectly through its effects (ordering) on conscious content, and on the other hand, any "observation of the unconscious," i.e., any operation of making unconscious content conscious, exerts initially uncontrollable reactions on the unconscious content itself (which, as known, in principle excludes the "exhaustion" of the unconscious by making its content conscious). The physicist will conclude by analogy that this uncontrollable reaction of the observing subject on the unconscious precisely limits the subjective nature of its reality and, simultaneously, confers upon it a certain subjectivity. Although further, the place of the "cut" between consciousness and the unconscious (at least to a certain extent) is left to the free choice of the experimenter in psychology, the existence of this cut remains a necessity. According to these data, the observed system would not only consist, from the psychological point of view, of physical objects but would also include the unconscious, while the role assigned to consciousness would be that of "means of observation."

Subsequently, it seems that the physicist and psychoanalyst traditions have dissociated. A lineage of researchers continued Pauli's ideas by seeking hidden variables in the formalisms of quantum mechanics (Bohm) or its algebra (Basil Hiley, Pylkkänen). Others, like physicist Bernard d'Espagnat, have utilised them to address the mind-body problem. Yet few authors to my knowledge have sought the formalisms employed in contemporary neuroscience.

While some neuroscientists claim to favour Spinozist monism[32], which historically is the first form of thought closest to that suggested by Pauli, this does not constitute a systematic and updated philosophical vision. It's uncertain whether the philosophy of contemporary sciences can easily adhere to the essentialism of Spinozism, or its causality concept inspired by Newtonian physics.

This is why it seems interesting, to test the viability of this hypothesis we introduced regarding the relationship between vivid consciousness and dead consciousness, to see if it is possible to exhibit structural symmetries between contemporary neuroscience theories and first-person experiential life. But before considering this path, which will constitute the second part of the thesis, it is necessary to provide a fourth and final analogy.

[32] Atlan is a vigorous proponent of Spinozism. See his work published in 2018.

12.4.4 Fourth Analogy: Dante's Hypersphere

A final analogy can be derived from more seriously exploring the idea that consciousness can be observed from two perspectives. Could we not imagine an object so complex in its entirety that the human mind is compelled to study several sections of it? To offer an easy metaphor, suppose we are reduced to observing a three-dimensional cylinder in a two-dimensional space. What would we see? The projection of the cylinder on a sheet of paper occasionally reveals the image of a sphere, occasionally that of a square, depending on the cylinder's orientation relative to the plane of the sheet. We call this square "matter" and name the circle "mind"—even when we do not just dismiss its existence outright. Yet, it is indeed the same object.

To extend the metaphor a notch further, let us now try to imagine what forms a four-dimensional object, such as a hypersphere, would take if projected into our three-dimensional space. The best way to visualise the hypersphere at minimal cost and with a hint of humour is arguably to glance at Dante's Inferno. The Italian poet, never stingy with illusions, represents hell inside the Earth and the celestial world outside it. Hence, the sphere of hell is encompassed within the sphere of paradise just as surely as Earth is within the Universe. If Dante starts from the Earth's centre, where Satan resides, and ascends to the surface in a straight line, where does he find himself? At the surface of both Hell and Paradise. This contact point between the two worlds coincides, for Dante, with the city of Jerusalem. If one departs from a random point in Paradise and progresses linearly in the same direction, where does one end up? In a three-dimensional world, one never leaves Paradise. After all, is Dante's Universe not infinite? But in the case of a hypersphere in the fourth dimension, one reaches the surface of Hell.

You understood me: the hypersphere is an infernal object! It must be seen as two spheres interconnected. Every time one exits one, one enters the second. Consequently, the concepts of "inside" and "outside" change meaning: the sphere of hell is within paradise; yet the sphere of paradise is itself in hell, and so forth. This unusual property is reminiscent of the relationship binding vivid consciousness to dead consciousness. Indeed, vivid consciousness seems "in" dead consciousness since phenomena unfold inside our heads. But dead consciousness is also "in" vivid consciousness, because in effect the world exists only in the consciousness we make of it.

I will argue that consciousness is precisely this: an object unrepresentable within our three-dimensional intuition. We cannot form a complete idea of it, but only metaphorical projections. One such projection is material space. The second is mental space. These projections are intertwined. They seem to be contained within each other, but they actually belong to the same object. I shall attempt to define a hybrid method, rational mysticism, capable of undertaking a systematic, albeit laminar and hypothetical, reconstruction

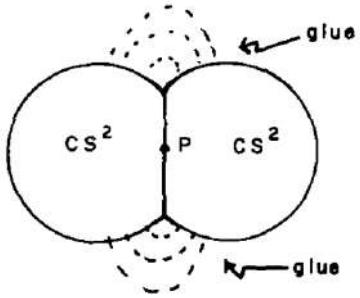


Figure 12.2: The hypersphere can be viewed as two spheres encompassed within one another. The only way to enter one is to reach the outer limit of the second. By always walking in the same direction, one thus cyclically moves from one sphere to the other.

of the underlying unity of the two dimensions.

Note that I used the term "in" with quotation marks. This precaution arises because it is often said that the mind is "in" the brain while I have continuously suggested, throughout these pages, the opposite. I leave to politicians the task of weather-vaning. For my part, I will continue to critique the idea that my head might contain a portion of consciousness. It is neither a bucket filled with magical water nor a lamp holding a genie, but the object of a far more complex mystery: "The fact that something as remarkable as a state of consciousness, said Huxley, occurs as the result of initiating a nervous tissue is as inexplicable as the Djinn's emergence once Aladdin rubbed the magic lamp."

What the metaphor of the hypersphere expresses is that the entirety of psychic space, taken as a whole, is contained within the entirety of material space, and conversely. So much so that the scientist will never find, by cutting neurons, anything resembling conscious life. On the contrary, to find the presence of the mind, one should reach the boundaries of matter. One might feel to approach this limit not by looking at substances inside the skull but at the brain's envelope itself, taken as an organic whole. Once embarked on this path, however, if one wants to be rigorous, one must extend the reasoning to the entirety of known matter, successively considering the skull's envelope, the human body's envelope, and then nature's envelope. This strategy is what makes Clark say that the mind is not *only* in the brain or the body but *also* in interaction with the environment:

> *It becomes increasingly difficult to say where the world ends and the individual begins*[33].

[33] Andy Clark, *Natural-born Cyborgs: Minds, Technologies, and the Future of Human*

This might give the aspiring anatomist the impression that consciousness is both everywhere and nowhere, immanent and transcendent, real and illusory. But what is surprising about that? Let us recall Dante's metaphor. It is by always walking straight that the traveller will emerge from the hell of atoms to enter the paradise of the mind. Yet it is just as impossible to exit matter by walking through space as it is impossible to think of the material world outside of thought.

So everything occurs as if conscious life occupied the infinitely distant boundaries of material space, and matter, in return, occupied the infinitely distant boundaries of conscious life. More precisely, I will say that the inner boundary of mental space, visible from this space though unattainable, is vivid consciousness. While the outer boundary of the mental, entirely invisible within the mental, is dead consciousness. This explains why we find nothing in our consciousness that resembles a brain, nerves, or ganglia. All we see of these inert objects are phenomena of consciousness: ideas, perceptions, representations, etc.

The visible boundary of matter, conversely, is dead consciousness. It is the one that becomes apparent at the surface of bodies, when we notice that removing certain parts of the brain or subjecting the organism to certain enzymes seems to affect the mind's functioning. The invisible boundary of matter, however, is vivid consciousness. It is the lived experience. It is what one fails to locate in matter. It is that invisible, intangible, immaterial soul dreamed by Swendenborg.

It is true, in four-dimensional space, that the mind is in matter and reciprocally. But in our human reality, we experience rather that the mind is neither truly "in" matter nor "outside" of it. At most, it is co-extensive with it. Body and mind are the two distinct facets of a single, unrepresentable reality. At most, we can offer parables, evocations, metaphors for it. The notions of "inside" and "outside" in this context are more akin to the ecstatic vocabulary of Meister Eckhart, Ibn 'Arabi, or the Upanishads than to our ordinary world experience. And though this vocabulary might impact as if clouds obscure the sun, I shall endeavour to show that clouds only bother those who remain on the ground of representational thought.

Intelligence, ed. Oxford University Press, USA, 2004. The translation is mine.

Key Takeaways from this Chapter

There exist only two ways to explain how a finite brain can execute intuitive functions:

- Either by modelling the brain with the formalism of transfinite machines;
- Or by resorting to an oracle (a black box).

These two approaches are unsatisfactory. I propose to defend a *perspectival mysterianism*. This consists of saying one can have a model of the brain and a model of the mind, but no model explaining the link between body and mind exists.

To justify this philosophical viewpoint, I make three hypotheses about consciousness:

1. Consciousness can be conceived in two ways: either *subjectively* (vivid consciousness) or *objectively* (dead consciousness).
2. There exists a logical paradox arising from the definition of these two concepts. This paradox is related to Russell's paradox.
3. Consciousness must therefore be considered from two perspectives: an "I" perspective and a "He" perspective.

In the next chapter, it will be shown that this vision of consciousness is compatible with the evolution of neurosciences.

Chapter 13

Critique of Encephalocentrism

> Not only our pleasures, joys, and laughter but also our sorrows, pains, and griefs come from the brain and only the brain.
>
> Hippocrates

In the realm of neuroscience, there's a tendency to conflate the concepts of consciousness, mind, and brain. This confusion is perhaps most evident in the French translations of scientific works. Gerald Edelman's "Bright Air, Brilliant Fire: On the Matter of the Mind" becomes "Biology of Consciousness[1]", while "Wider Than the Sky: The Phenomenal Gift of Consciousness" is translated to "More Vast Than the Sky: A General Theory of the Brain[2]". These semantic shifts reveal a blend of explicitly materialist metaphysics with implicit references to first-person phenomenological experiences.

In the second hypothesis on consciousness (section 12.3), I described the externalist chiasm as the tendency to confuse a purely functional *description* of the brain—filled with myelin, synapses, axons, neurotransmitters, and electrical potentials—with a truly convincing *explanation* of consciousness. Within this chiasm, it seems as if the *easy problem* of consciousness, which involves understanding the brain's operation in scientific terms, is deftly switched for the hard problem of consciousness — which seeks to grasp why I experience sensations and how there exists a seemingly magical correspondence between my organism's responses and the subjective experience of thinking, suffering, enjoying, and living the world in a deeply personal way.

[1] *Biologie de la Conscience* en français.
[2] *Plus vaste que le ciel. Une théorie générale du cerveau* en français.

So far, I have discussed these metaphysical chiasms as if they were established truths without explaining their genesis or relevance to modern theories. Hence, two immediate questions arise: (1) is the externalist chiasm present in neuroscience? and (2) what are its historical and epistemological origins? In particular, we might ask whether it has always been accepted that consciousness is a property of the brain. If not, when did this idea emerge?

To illuminate these questions, this section will begin with a brief critical history of the idea that consciousness resides in the brain. This will provide an opportunity to highlight some foundational principles of contemporary neuroscience, such as lesion methodology (study of brain-damaged patients), the principle of functional segregation (each part of the brain has a role), and neural plasticity (brain parts can be reallocated). It will also raise the question of how the brain's parts relate to the mind itself and how the sensory organs relate to the organism as a whole.

In the chapters that follow, I will focus more specifically on tracing the presence of metaphysical chiasms in today's theories of consciousness. By "prevailing theories", I do not necessarily mean those with unanimous scientific approval or even current usage. Rather, I am interested in theories that have significantly influenced the way the problem of consciousness is approached from a neuroscientific perspective. This influence often relies more on the nature of the epistemological principles employed than on the technical and scientific details of the ideas themselves.

A chapter titled "The Darwinian Theory of Gerald Edelman" will explore Gerald Edelman's theory of neuronal group selection. This Nobel laureate in physiology (1972) critiques cognitivism and introduces a deeply emergentist view of consciousness, seeing it as a property not of individual neurons or groups of neurons, but of the brain's overall activity. I will attempt to demonstrate the advantages of his non-computational approach (the mind is not a computer) and his non-dualist perspective (the mind is not separate from matter), while highlighting an externalist chiasm. The risk of discussing causal powers and the lack of a genuinely phenomenological analysis of what it means to "be conscious" will be under scrutiny.

Another chapter, simply titled "Integrated Information Theory", will focus on the theory by Tononi and colleagues, extending Edelman's work with inspiration from Shannon's information theory. I will suggest it is one of the few neuroscientific theories to seriously engage with the truly phenomenological aspect of the consciousness question. It endeavours to draw a structural analogy between the "axioms" of consciousness, as revealed in the introspective "I think", and the fundamental operating principles of the brain. Nonetheless, a contradiction in the description of lived consciousness will be noted. Additionally, I'll critique the notion that consciousness is something immediate and representational. I will also challenge the idea that consciousness is merely information processing.

At the conclusion of this journey into the enticing yet potentially reductive world of neuroscience, I aim to highlight a number of essential epistemic principles, according to my working hypotheses, for building a monistic and bi-perspective theory of consciousness, where body and matter are two facets of the same ontology. It will become apparent that the key elements lie in mereology (the articulation of whole and parts), structural analogy (the laws of the mind and the laws of matter align), integration (unity of representations in the act of consciousness), and differentiation (consciousness focuses on changes).

13.1 Is Consciousness in the Brain?

The brain's regulatory role is a fact known as old as humanity, or at least since humanity's historic awakening: the Edwin Smith papyrus, dating from the VIth or XVIIth dynasty, records no fewer than 48 lesions, mostly war injuries, about a quarter involving the skull or spine. It contains the first reference to the word "brain," along with passages indicating an understanding of the sympathetic and parasympathetic nervous systems:

> *"If you examine a man with a dislocated neck vertebra, and find he is unaware of his arms and legs, while his phallus is erect, and urine drips from him unknowingly, while his flesh has absorbed air, and his eyes are bloodshot: it is a dislocation of a neck vertebra joining his spine that has rendered him unaware of his limbs."*

It is one challenge to pragmatically describe how the body functions, but quite another to grasp the subtle link between corporeality and the origin of our sensory capabilities. Despite early surgical insights, numerous debates have occurred about the brain's role in sensory, emotional, and cognitive functions. In the nascent stages of medical science, often credited to Hippocrates, two visions stood opposed: encephalocentrism and cardiocentrism.

While the undisputed father of medicine, visionary as he was, acknowledged more to the brain than mere body temperature regulation, this was not a view shared by all his contemporaries. Aristotle, and to a lesser extent Praxagoras, championed the heart as the seat of sensation. It was not until Herophilus, Erasistratus, and later Galen that encephalocentric views gained precision and prominence, only truly triumphing in the 16th and 17th centuries, under the works of Vesalius or Thomas Willis. Even after the idea of cerebral dominance entered medical discourse, the ventricular theory, which posited the brain's empty spaces played a key role, maintained firm footing.

Today, it is hard to refute the correlation between mental functions and the organism's comprehensive operation: aren't heartbeat, galvanic skin

response, respiratory rate, body temperature, and pupil diameter reliable indicators of our mental and emotional states? If not, why are they still used in lie detectors? Though brain damage seemingly results in the cessation of conscious activity — without knowing what happens to consciousness post-mortem — one could argue the cessation of the heartbeat has the same effect. Why, then, should the brain be given a predominant role in the presumed production of consciousness?

Neuroscience doesn't offer a definitive answer, but it continues to accumulate significant evidence that various brain parts are closely correlated with specific cognitive functions that are implicitly regarded as integral to consciousness. In this regard, Thomas Willis can be considered the first true pioneer of a brain theory that appreciates cortical zones, with notable distinctions such as white matter and grey matter. In his treatise *Cerebri Anatomy*, he categorizes cranial nerves and associates certain cognitive functions with the striatum or the corpus callosum. Voluntary movements, for instance, were located in the cortex, while involuntary movements were attributed to the cerebellum [3]. Modern neuroanatomy, honouring his contributions, names the anastomotic polygonal structure at the brain's base after him.

From this point, the central question shifted. It was no longer about which, the heart or the brain, was the seat of thought, but how specific intelligence functions relate to well-known brain areas, almost like a neurophysiological periodic table. This shift began with anatomists Franz Joseph Gall (1758-1828) and Johann Caspar Spurzheim (1776-1832), the firm advocates of phrenology, known as the "theory of bumps" because it claimed mental functions were linked to the brain's shape. Despite a promising investigative future, and support from Auguste Comte[4], this theory foundered against the holistic view of brain function.

The main scientific critique of phrenology, aside from methodological issues[5], stems from Pierre Jean-Marie Flourens' (1794-1897) experiments with pigeons. It was observed that these animals gradually recovered from induced brain lesions. The critical factor seemed not the location but the amount of brain matter removed. This eloquently suggests brain plasticity, without negating the phenomenology or the localisationist intuition underlying its conclusions. And rightly so: the concept of functional segregation remains central to the neuroanatomical approach, as evidenced by two key subsequent discoveries.

First, Hughlings Jackson's (1835-1911) observations on epileptic patients revealed that seizures followed a specific progression through the body, as

[3] 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.

[4] Cf. Comte Auguste, *Système de politique positive*, Hermann, 2022.

[5] See for example Hegel's chapter on phenomenology.

if each localised brain excitation corresponded almost bijectively to a body part. He thereby postulated functional somatotopy, foreshadowing Wilder Penfield's homunculus of the precentral gyrus and the sensory homunculus of the somesthetic cortex[6].

Similarly, the famous case of Monsieur Leborgne, known as "Tan-Tan" due to the only utterance he could produce, allowed Paul Broca (1824-1880) to pinpoint the language production area at the left third frontal convolution, now known as Broca's area. Not content with laying the groundwork for neuropsychology, this was closely followed by Carl Wernicke's (1848-1804) discovery of a brain area for language comprehension. From his clinical observations, Wernicke localized this function in the left temporal lobe and predicted other aphasia types, like global or conduction aphasia, pursuing the dream of a Mendeleev-style brain map[7].

13.2 Functional Segregation and Neural Plasticity

Today, although holistic approaches have largely faded, the notions of functional segregation and neural plasticity seem to coalesce coherently in brain representations. I refer to representations in the plural because early neuroanatomy, through natural expansion, has spawned a plethora of new sciences: among them, neurohistology, neuroendocrinology, neurophysiology, neurology, neuropsychology, electrophysiology, and neuropharmacology. Despite the diversity of these neurobiological fields, they seem to cluster into four main approaches: top-down, bottom-up, structure-function correlation, and organised systems[8].

The shared hypothesis of all these approaches is that it's possible to explain an individual's intimate mental life, viewed from the "I" perspective, through an apt theory of the brain. It is as if the age-old debate between encephalocentrism and cardiocentrism merely changed in nature, keeping its original concerns relevant. While all agree that expressions like "learning by heart", "having a broken heart", or "having one's heart on one's sleeve" are metaphorical, one might question the literal meaning of expressions like "being the brains of a plot", "brain drain", or "rack one's brains"? Without falling into the cliché of quip-tossers who might claim "falling in love is chemical[9]", there is a tendency to view the brain analogous to what the ancients attributed to the human heart: the centre of thought and emotions.

[6]Cf. PENFIELD W, BOLDREY E, *Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation*, Brain, 1937 and PENFIELD W, RASMUSSEN T., The cerebral cortex of man, New York: Macmillan, 1950

[7]WERNICKE C, *Grundriss der Psychiatrie in klinischen Vorlesungen*, 1900.

[8]This categorisation is borrowed from CYRIL MONIER, Neuroscience within the Cognitive Sciences: A Review of the Literature Advocating for a Spinozist Future of Neuroscience, 2019.

[9]VINCENT L., How Does One Fall in Love?, Paris, Odile Jacob, 2004.

It's crucial at this point to differentiate genuine neuroscientific practice from its portrayal by critics. It would be too facile, from our standpoint, to attack an easy target[10]. A portion of the so-called "neuroskeptic" literature mischaracterises neuroscientists as naive philosophers, staunch reductionists who trumpet grandiose claims about the exalted status of science[11]. Yet, upon examining practice, it appears that the "bottom-up" approach, which seeks to elucidate higher-order cognitive functions from neurons, isn't the dominant strategy in neuroscience. According to Daniel Stoljar and Ian Gold, the majority of neuroscientists align with the emergent systems paradigm[12]. Hence, they might endorse a form of weak reductionism or even profess metaphysical ignorance.

It might be insightful to accept that reductionist thinking in neuroscience, which some refer to as the neuron doctrine[13], can be interpreted in two distinct ways. Following Daniel Stoljar and Ian Gold's article, the claim that only cognitive neuroscience can account for the mind's workings might be understood in two different manners:

- In a trivial sense: considering "cognitive neuroscience" in its broadest definition. It's this interdisciplinary approach traditionally positioned at the crossroads of biology and psychology, prescribing a methodology and a goal, but no specific theory. The only underlying assumptions, according to Stoljar and Gold, are threefold: an adherence to materialism, since the mind isn't another substance from the physical world, though emergence or supervenience may be present; an acceptance that the mind can be accurately described by scientific theory, akin to a form of naturalism; and the idea that biology alone is insufficient, necessitating possible recourse to psychological concepts. Psychology might propose functional explanations while biology deals with mechanisms.

- On a radical level: by substituting "cognitive neuroscience" with "biological neuroscience". This makes the doctrine much more ambitious, implying concepts in linguistics or cognitive psychology are mere simplified expressions of neurobiological phenomena. A working memory theory discussing "phonological loops", "visuospatial sketchpad",

[10] Cf. RAMUS F., *The Convenient Straw Man*, 2014.

[11] An example of such unfortunate assertions can be found in Crick's work: "You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules. You are nothing but a pack of neurons."(CRICK, F., *The Astonishing Hypothesis*, New York, Charles Scribner's Sons, 1994).

[12] GOLD, I, STOLJAR D, *A neuron doctrine in the philosophy of neuroscience*, Behavioral and Brain Sciences, 22(5), 809-30; discussion 831-69.

[13] *Ibidem.*

> or "episodic buffer" would thus be a lay explanation, or a stepping-stone, towards truly understanding the phenomenon[14]. In this radical vision, more is claimed than the mere scientific descriptibility of the mind, albeit at a cost to plausibility.

This distinction has sparked numerous debates, which we will dissect in due course. What we wish to highlight for now, as a preliminary, is that even in the trivial conception of the neuron doctrine, the brain remains pivotal. It's not just central because it's the most important human organ, without which humanity can't be conceived; it's central because explanatory relationships seem plausible between the brain's biological mechanisms and the psychic experiences related to direct consciousness experience. Here, we consider the audacity of neuroscience to bridge the gap between third-person viewpoints and first-person perspectives. If indeed we wish to champion, as announced in the introduction, a dual-aspect monism, it is essential to address this substantive objection from science, asserting the human brain as a bridge between these pronominal viewpoints.

In their original articles, Stoljar and Gold consider only three potential alternatives to the materialistic naturalism underlying the trivial neuron doctrine. The first involves rejecting the hypothesis that (a) the mind is a biological phenomenon. They cite dualism as an example, which continues to garner considerable support today. The second approach involves denying the hypothesis that (b) the mind is the natural sciences' study object. They illustrate this possibility with social constructivism theories, although they note dualism is also a strong candidate. The third strategy involves rejecting the idea that (c) biology and psychology are sufficient to explain the human mind. A theory fitting this frame might be the programme of strong artificial intelligence, in its purist form, considering a theory of mind in "abstract terms[15] ", which would not require the support of these two disciplines.

13.3 Neuroscientific Spinozism

We propose a fourth path. We believe it is logically viable and aligns with neuroscientists' practices. This approach argues against the notions that (a') the mind can only be satisfactorily comprehended when viewed solely as a biological phenomenon, and (b') the mind can solely be studied by the

[14] I'm thinking here particularly of Kim's article (1992, 1999), which systematised this issue, sparking a whole range of emergentist theories. For a more extensive taxonomy, see also Sartenaer Olivier (2015).

[15] These terms are used by the authors: "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).

natural sciences. The emphasis on "only", highlighted here, does not hint at hidden dualism. It refers instead to the metaphor introduced earlier: just as a quantum object can be seen as both a particle and a wave (Bohr's principle of complementarity), we advocate the idea that the mind can be understood both phenomenally and materially[16].

Far from imagining the mind as a volatile substance trapped in matter, like a ghost in a machine, or reducing the mind solely to matter, we suggest mind and matter are two complementary, albeit mutually exclusive, perspectives for representing the world — and ourselves within it. If the "idealist" perspective originates from the first-person viewpoint (the "I" perspective), the "materialist" perspective originates from the third-person viewpoint (the "It" perspective). The former rests on what Husserl called the transcendental phenomenological regression. The latter is rooted in the surrounding world's stability and our pragmatic ability to predict it. As it stands, proving the existence of an external reality to our phenomenal experience, with a definitive argument, or explaining the science miracle without invoking a tangible world is difficult. These representations appear viable.

That does not mean two distinct substances exist. Instead, the world is conceivable both ways, but not simultaneously. Dualism's central problem is accounting for the link between matter and mind[17]. This issue doesn't arise under our hypothesis, as dualism isn't ontologically implied. The two modalities are merely human perspectives on a single substance where the information partially coincides. Thus, from the "It" perspective, almost everything can be materially explained, while from the "I" perspective, almost everything can be explained within consciousness. Difficulties arise only at the intersecting points between these perspectives, at the boundaries of their explanatory spheres. This relates to questions like how a matter aggregate can have intimate life, a property sometimes called qualia or the phenomenal "I". Similarly, within a first-person approach, questions about the strong links between cognitive functions, like language, and certain brain areas (Broca's or Wernicke's area) arise.

In both cases, the issue results from combining vocabulary specific to these perspectives, essentially creating a chiasm. Whether this contradiction stems from human perspectives' inability to apprehend reality fully, or reflects something fundamental in nature, is speculative. This will be considered in the thesis's second part, alongside various other objections. Currently, we're merely outlining the methodological hypothesis that certain elements in the first-person viewpoint, methodologically, can't be reduced to the third-person viewpoint; and vice versa.

[16]Note: this metaphor concerns our access to knowledge, not the nature of neuronal functioning or consciousness itself. Nothing is claimed about quantum physics' relevance to addressing certain philosophy of mind issues.

[17]I'm thinking particularly of Kim's article (1992, 1999), which systematised this issue well, leading to various emergentist theories. See also Sartenaer Olivier (2015).

It should be noted, before proceeding, that these two viewpoints, “I” and “It”, share a common trait: self-consistency. Although incomplete, each suffices independently and doesn't require the other for conceptualization. Certain phenomena emerge that neither can satisfactorily explain and become stumbling blocks, but only when attempting cross-explanation. Thus, explaining the mind from matter (reification) or matter from the mind (idealisation) is futile. Statements like "this neuronal group's activity contributes to consciousness" contradict in terms unless "consciousness" is viewed purely as an objective material manifestation. What do we mean by that? Indeed, an experiment can involve a subject with an encephalogram, studying relations between measured signals and self-reported conscious perceptions. It is also possible to study effects on self-reported perceptions after specific brain lesions. Such an experiment establishes a definitive correlation between neuronal activity and psychological experience. It reinforces the brain's necessity for thought, not as a cause, but that neuronal activity and self-reported conscious perception (a behavioural fact) occur concurrently. Crossing into linking matter with consciousness itself, in its lived essence, is akin to crossing Hume's induction gap from apodictic truth to plausibility. Correlation suggests underlying reality but transcends scientific validity.

When considering mind as something extended, we lose sight of its essential property. It becomes a tangible, observer-independent object, not a lived experience. Conversely, viewing matter as an act of consciousness, akin to subjective idealism, reduces it to its phenomenal manifestation. There's no more sense in explaining how thought can emerge from neurons than questioning the existence of a thinking substance within cells. Does this mean exploring the link between them is futile? It depends entirely on the approach. If indeed thought and matter are two facets of the same reality, it makes sense to believe this reality follows identical laws. It might thus be possible to demonstrate a common structure between the laws of mind and matter. This structure would clarify how neuronal inputs (the “it” perspective) interconnect and how phenomenological experiences (the “I” perspective) relate. Such a structure wouldn't constitute proof but is necessary for the hypothesis's viability. This is why isolating the epistemic principles underpinning mind formation in neuroscience is crucial to determine compatibility with those discovered in a first-person investigation and potentially falsify our hypothesis.

Although this hypothesis of dual modality may seem foreign to neuroscience, epistemically speaking, it is indeed endorsed by several eminent field professionals. Its inspiration is closely linked to Spinozism:

> *"The mind does not emerge from matter as if it were caused by it, nor does it cause the body of which it is the idea, or any other. There is an emergence of the brain's material complexity*

> *from the simpler elements — neurons and their connexions — that comprise it. And there is an emergence of the mind, the idea of the brain (or, more precisely, of the body) from the ideas of the neurons, being understood [...] that the idea of a neuron, like that of an electron or a stone, is nothing other than its law of organisation and functioning. This is simpler, although also pertaining to the attribute Thought, than the idea of the human body, i.e., the human mind. The properties of the latter emerge from the interactions among the ideas of the body's constituents. The human mind and body thus both have emergent properties from their constituents. But they are both, as well as their respective constituents, two aspects of the same thing viewed under different attributes." (Atlan, 2018).*

The strategy in the upcoming chapters is thus twofold. Firstly, to show that in neuroscientific works, there exists a blurred line between the first-person perspective and the third-person perspective. Despite certain precautions, something more is often implied about the brain than what is generally said about other organs. Beyond being seen as a mere necessary condition of thought, or phenomena coinciding with the subjects' self-reported conscious properties, neuronal models typically attempt to capture, at least partially, the subjective aspect of conscious experience.

Secondly, the aim is to isolate the epistemological principles that facilitate the mind's understanding and demonstrate their compatibility with the dual-aspect monism, or at least the neuroscientific Spinozism, that I aim to defend in this thesis. The conclusions drawn will guide the subsequent development of a mereological theory of consciousness.

Key Takeaways from this Chapter

In scientific practice, there's often a blurring between the "I" perspective (subjective consciousness) and the "It" perspective (objective consciousness). This blurring is what I refer to as:

- an *externalist metaphysical chiasm*, when the "I" is reduced to the "It."
- an *internalist metaphysical chiasm*, when the "It" is reduced to the "I."

These chiasms can be avoided if, instead of trying to reduce lived consciousness to a biological theory of the brain, we accept that matter and consciousness are two perspectives on the same reality.

I've demonstrated that this dual-aspect monism is aligned with neuroscientific Spinozism and is supported by many in the scientific community. I also suggest that, if consciousness and matter are two faces of the same reality, they must share a common structure.

This implies a potential analogy between the principles governing consciousness (considered as an introspective "I") and those governing the brain (viewed as an observational "It").

The following chapters will explore two neuroscientific brain theories, those of Edelman and Tononi.

Chapter 14

Gerald Edelman's Darwinian Theory

> Nothing in biology makes sense
> except in the light of evolution.
>
> Theodosius Dobzhansky

It's not uncommon for distinguished scientists to suddenly feel compelled to speak on a vast array of topics once they've been awarded the Nobel Prize—and are encouraged by the media to do so. Whether it is politics, economics, happiness, extraterrestrial life, or consciousness, it seems that the practice of natural sciences, for reasons not unlike Saint Thomas Aquinas's theory of revealed grace, predisposes the expert to seize, without warning, the entire celestial sphere of Platonic ideas.

Such phenomena usually provoke a healthy skepticism, yet when these shifts are confined to a specific topic, such as moving from immunology to consciousness, or from biology to epistemology, and when undertaken by a scholar like Gerald Edelman, who blends a talent for popularisation with philosophical circumspection, they can lead to intriguing research avenues[1]. In this era of unprecedented acceleration, perhaps science needs nothing more than a comprehensive, synthetic, multifactorial approach, capable of transferring knowledge from one domain to another through an artful interplay of communicating vessels, thereby uniting, converging, revealing—and even dusting off—the traditional disciplines.

[1] In his address to the Nobel Committee upon receiving his award in 1972, Gerald Edelman did not hesitate to explicitly articulate the idea that: "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 became renowned for his pioneering research on the structure of immunoglobulins, or B lymphocytes, for which he received the Nobel Prize in Medicine in 1972 alongside his colleague Rodney Porter. The object of their success? They managed to determine amino acid sequences of monoclonal antibody chains from patients suffering from an immunoproliferative syndrome. Not content with this, he sought to answer how a general structure of lymphocytes could give rise to such a variety of antibodies, tailored to each antigen molecule. At the same time, he developed the bold concept that multiple genes control the synthesis of antibody molecules, some responsible for the constant part of the molecule and others for the variable part. This pioneering work would later be confirmed by molecular biology.

After winning the Nobel Prize in 1972, G. Edelman attempted to transpose the results of his immunological theories to the study of the brain and consciousness: "These mechanisms form the basis of what I have called the theory of neuronal group selection [...] All processes at work in the development and morphogenesis of the brain have a certain statistical side: neurogenesis, neuronal migration, adhesion that regulates the growth of neuronal extremities, programmed neuronal death, connexion formation [...] These variations combine to create an absolutely extraordinary diversity of structures, resulting in no two individuals, not even twins, ever having the same brain. This is developmental selection, the first pillar of my theory. Subsequently, the wiring between neuronal groups is modified based on interactions with the environment, especially sensory[2]..."

In this chapter, I aim to provide an overview of Edelman's theory of consciousness, addressing both its scientific and epistemological aspects. I shall then attempt to demonstrate the presence of an externalist metaphysical chiasm, involving a blend of observational terms, referring to objective phenomena of space, with introspective terms that relate to a subject's experience of the world. I will conclude by offering some pathways toward developing a biperspectivist theory of consciousness.

14.1 Edelman's Critique of Dualism

Before delving into Edelman's theory of consciousness, I would like to identify the epistemological intuitions that guided his approach. Was Edelman a proponent of materialist reductionism or dualism? Did he believe in two distinct substances (body and mind), or did he think consciousness could somehow partially reduce to a set of physicochemical laws? I will begin by discussing the objections Edelman raises against the idea of a substance separate from matter, and then I will examine, contrary to this opinion, the

[2]Cited from the French Academy of Sciences, https://www.academie-sciences.fr/fr/In-memoriam/gerald-maurice-edelman.html

objections he raises against the notion that consciousness functions like a computer. This will pave the way for a third approach, known as emergentism, in which consciousness, though supported by material processes, does not reduce purely and simply to neuronal operations.

In his work *Second Nature: Brain Science and Human Knowledge*, Gerald Edelman immediately reveals his metaphysical convictions. He criticises Cartesian dualism, which distinguishes between the *res extensa* (matter) and the *res cogitans* (thought), for having delayed the scientific study of the human mind.

> *One factor affecting the scientific approach to consciousness can be attributed to the very influential thought of René Descartes. Some time after Galileo, Descartes withdrew the mind from nature. He did this through thought alone, concluding that there existed two distinct substances: the res cogitan, extended things falling under physics, and the res cogitan, thinking things having no extension in space and not accessible to physics*[3].

The *res cogitan* defended by Descartes would, according to Edelman, be a substance refractory to the scientist's lens. Since this order of reality is accessible only via the perspective of "I think", it cannot be the subject of established objectivity. This leads him to claim that the human mind has been removed from nature. His dismay is understandable. In Descartes' view, consciousness is no more accessible to scientific observation than distant planets are visible to the naked eye. The reality of the mental is relegated to the invisible world of intimate phenomena.

While it is understandable that Edelman attacks the *res cogitan*, which he sees as a residue of the concept of the soul, it is nevertheless legitimate to question to what extent consciousness has resisted scientific theories not due to prejudice, but because this consciousness does not submit to the experimental method? Science, taking an "It" point of view, cannot account for the genuinely phenomenological or subjective dimension of the mental world. Descartes' merit lies not in having retarded science but in having sought to unite the genuinely subjective and objective dimensions of thought.

The way he proceeds in the epistemological domain is certainly debatable. Nevertheless, Descartes relies on scientifically established notions of his time: I refer, on one hand, to animal spirits and, on the other, to the pineal gland. The latter, located at the entrance of the Sylvian canal, is today known to produce melatonin, involved in regulating sleep. Although it plays no particular role in modern neuroscientific theories of consciousness, Descartes saw in it a symmetry argument: it was the only known brain organ devoid of duplicate in the right and left hemispheres. It occupies, like the nose in the middle of the face—or the throne in the middle of the skull—a

[3] Edelman G., *Second Nature*, Yale University Press

central position. Unlike the eyes and ears, which provide two distinct images or sounds of a single object, the gland unifies these diverse impressions into a single perception. This supports its integrative function specific to consciousness[4].

Descartes' ideas, in some anatomical aspects, represent a regression from the established knowledge since Galen's work. For instance, when he claims the pineal gland is suspended in the ventricle, not outside; or that this gland is composed of a volatile, not liquid, substance. Scientific errors indeed. Thus, the recourse to the notion of the soul indeed contributed to popularising a somewhat magical view of cerebral anatomy. In this, Edelman is correct. However, it is essential to distinguish, giving Descartes his due, between the content of ideas and the intention underlying the approach.

Ultimately, Edelman's persistence in attacking dualism may not rest so much on Descartes' scientific errors as on a possible confusion with what Chalmers called the easy problem and the hard problem of consciousness? It's relatively "easy", at least in principle, to build a theory explaining how the brain integrates sensory information to produce evolutionarily adapted reactions to environmental constraints. It's also "easy" to relate brain mechanisms to self-reported phenomena of pain, thought, or imagination. However, it is challenging to explain how being an "I" thinking the world is produced by the brain:

> *It is of utmost importance to take consciousness seriously. The simplest way to develop a "theory" of consciousness is to deny its existence or to redefine the phenomenon needing explanation as something it is not. This usually leads to an elegant theory, but the problem does not go away*[5]

14.2 Edelman's Critique of Computationalism

Despite his critique of dualism, Edelman does not defend a blind physicalist reductionism. He does not envision the brain as a computer tasked with processing incidental information to produce, as a side effect, mental states. He emphasises, instead, that the mind is embodied and the body itself is integrated into a broader environment. Whenever he speaks of a brain property, he clarifies it as a simplification to designate the entirety of

[4] This theory of Descartes on the pineal gland is not an innovation per se. In Galen, we find affirmations that the ventricles connected to this gland were filled with *pneuma*, a volatile substance considered by the ancient Greeks as synonymous with the soul. In antiquity, different cerebellar appendices were attributed a role in regulating the pneuma.

[5] Translation by me: "*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*.

signals exchanged between organs, wherein the brain holds a central role as a regulator. The exchanges between body and brain, in turn, are conditioned by a certain relation to the environment, forming what we call an eco-niche[6]. In this respect, Edelman situates his thought within the enactivist paradigm previously discussed.

Beyond the emphasis on the body and environment, Edelman explicitly critiques the reductionist vision (*reductio ad computatum*[7]) where the computer serves as a model for brain function, and this for three fundamental reasons. Firstly, Edelman reminds us, a computer relies on arithmetic and symbolic language. It manipulates objects in short time intervals, regulated by an internal clock. There are predefined time units, during which an instruction like: "subtract quantity a from variable b", "assign value a to variable b", etc., can be read. However, Edelman argues that the brain isn't governed by logical rules but operates, as I will subsequently show, according to a function emerging from the selective association of neuronal groups.

Besides, a computer requires unambiguous input to function correctly. It doesn't work with analog magnitudes but rather natural integers. Yet the signals handled by the brain are not well-organised: far from resembling characters on a tape, the physical world, the tangible world of everyday life, doesn't present intrinsic structure. The whole challenge of artificial intelligence lies in organising sensory data into categories convenient for manipulation. That is why the first renowned instance of neural networks, a cornerstone for today's educators, consisted of reading hand-written numbers. The task was to take a vast array of complex visual characters, composed of dots, lines, scribbles, more or less thick strokes, reduce them to identifiable pixels, and then associate these pixel groups with numbers 0, 1, 2,... In Kantian terms, one might say the specificity of the brain is to subsume sensible content, this rhapsody of images, sounds, colours, under a ready-to-use concept.

Edelman's second critique highlights that structuring neuronal networks follows a process unique to each individual. "There are no two identical brains, not even those of twins. This is largely due to the fact that during development and the establishment of neuroanatomy, neurons that fire together become connected[8]." It is challenging to see how exactly this critique represents a rigorous argument against computers, given that neural networks can also take on unique value sets during their training. Edelman seems to refute the idea that neuroanatomical development of the brain can be likened to the "intelligent learning" familiar to programmers. These pro-

[6] Gerald M. Edelman *Second Nature: Brain Science and Human Knowledge*, p.25

[7] This expression is mine. It ties in with what was said in the introduction of the first section.

[8] Gerald M. Edelman, *Second Nature: Brain Science and Human Knowledge*, Yale University Press, New Haven and London, p.21

grammers indeed define a set of evolution rules regulated by a logical clock, whereas the brain results from a selective process comparable to immunology, which retains only effective neuronal networks. In Edelman's view, evolutionary programming, though inspired by biology, doesn't genuinely capture the function of the cerebral organ.

Edelman's third argument is that humans do not possess enough genes to completely code brain function. It's straightforward to demonstrate that the number of possible synaptic connexions far exceeds the number of genetic combinations, implying that part of learning is not pre-coded in the organism but relies on emergent phenomena from environmental contact. While interesting, this argument seems again overly simplified. It draws an analogy between genetic code and computer code, concluding that if the brain were a computer, its function would need to be perfectly described by its genetic code. This implicitly presupposes that a computer's function is fully described by its code. However, concerning artificial neural networks, the code only defines learning rules from which key variables may converge toward certain value sets. The final disposition of artificial neural networks emerges independently of the programmer's initial intent—thus, predicting it based solely on the initial code would be futile. Edelman's third argument either boils down to the first—that no logical rules in the brain are comparable to a computer's—or posits that the genetic code is too weak to define even all synaptic variables involved in the brain's neuroanatomical formation.

To understand Edelman's persistence in challenging logical clocks and genetic predetermination, some contextual elements are helpful. In the 1970s, the period when Edelman was developing his initial brain theories, there was a prevailing idea, firstly, that genes coded brain behaviour; and secondly, that they did so centrally. It must be noted that Watson and Crick's DNA discovery, still quite recent, had captivated dominant scientific narratives, making the gene the cornerstone of a comprehensive mechanical explanation of human behaviour. This "genetic machinery", as described by Thierry Magnin[9], was expected to explain intelligence development, personality, etc. Today, epigenetics popularises the notion that some genes, while present in the organism, will be expressed— or not— based on the subject's experiences, psychological experiences (stress, mental exercise, joy, spiritual practices...) and complex environmental interactions. We've shifted from a centralised vision of the human machine to one emerging from interaction with numerous components.

A striking example of epigenetics, and worthy of inclusion in an authentic crime series, is the case of James Fallon. This University of California neuroscientist sought throughout his life to characterise cerebral and genetic

[9]This phrase, alongside others of similar nature, can be found in his oral presentations or his article *Human Measure in Post and Transhumanism*.

components of psychopathy, until, by chance, he stumbled upon an MRI image exhibiting all the markers he identified in his work. Imagine his shock when realising it was his own brain's MRI! James Fallon had the specific gene for psychopathy, but his favourable life context hadn't allowed this gene to manifest—although news of this genetic peculiarity didn't particularly surprise his entourage[10].

These considerations lead us to the following question: if the brain's configuration isn't fully described by genetic code, how does the brain develop according to Edelman? This requires explaining his neurodevelopmental theory known as "Neuronal Group Selection Theory" or "Neuronal Darwinism."

14.3 Theory of Neuronal Group Selection (TNGS)

Given the preceding discussions, it becomes clear that according to Edelman, the theory of consciousness must address several constraints, including the diversity and regularity of responses provided by the brain, all without presupposing the existence of a clock or logical rules that equate the mind to a computer. Since, as noted, genetic material alone does not suffice to prescribe the brain's overall configuration, Edelman, to answer this dilemma, draws inspiration from the theory that garnered him the Nobel Prize in Medicine with Rodney Porter: the immunological theory describing antibody function.

It is widely known that an organism, after exposure to a specific viral strain, becomes adept at responding to related threats. From an immunologist's perspective, this suggests that antibodies can recognize foreign molecules, remember this information, and neutralise them upon subsequent encounters. A naive theory might suggest that antibodies, much like modeling clay, can envelop encountered antigens and retain their contours even in absence. A lymphocyte population, exposed to a molecule, thus mechanically develops the necessary form to identify that particular molecule. This conception likens lymphocyte memory to a clay tablet where necessary information is inscribed. However, this conception is incorrect.

In 1972, Edelman developed a theory based on a Darwinian mechanism. Instead of equipping antibodies with a shape-modulation ability, it was proposed that each lymphocyte in the body possesses a distinct type of antibody on its surface, capable of recognising and binding to a specific antigen. With nearly a hundred billion lymphocytes in the body, there exists a vast stock of diversified antibodies. Upon introducing a foreign molecule into the bloodstream, random collisions between this molecule and lymphocytes eventually result in an encounter with an antibody capable of binding

[10] This anecdote is recounted by the main party in an essay titled *The Psychopath Inside: A Neuroscientist's Personal Journey Into the Dark Side of the Brain.*

to it. Thus, a shape change assumption is unnecessary. The lymphocyte bearing the appropriate antibody duplicates itself to proliferate. If the same molecule is reintroduced into the organism later, a much larger proportion of antibodies will then be capable of binding to, identifying, and neutralising it. This mechanism elucidates immunological learning's foundation, notably underpinning vaccines.

The selection mechanism relies on three ingredients. Firstly, what is termed a diversity generator, responsible for the initial, varied antibody population. Then, there must be environmental interaction, ensuring each antibody has the capacity, or lack thereof, to bind to a foreign molecule. Finally, a mechanism must ensure prolific replication of antibodies having found their target. In general terms, these three components are referred to as diversification, stimulation, and proliferation. The remaining question is whether these three elements are visible in the formation of nerve cells?

Edelman affirms they are. He argues that nerve cell diversification arises due to the significant microscopic variation observed in neuroanatomical development. To begin with, it's worth noting that nearly two hundred distinct neuron types exist, characterised by their excitability and conductivity. Excitability indicates the input potential needed for a neuron to trigger a neural impulse, while conductivity depends on neurite structures, particularly axon length and number of dendrites (see Figure 14.1). For instance, a pyramidal neuron in the motor cortex has a very long axon ensuring communication with nervous system regions. Conversely, Golgi type II neurons have very short axons with few nerve endings, found in the cerebral cortex, cerebellar cortex, and retina.

Not only do numerous diverse neurons exist, but connexions among them simultaneously allow considerable evolutionary variation. A simplified brain analogy involves neuronal population presenting random excitations from foetal development to advanced evolutionary stages. Under Hebb's rule, neurons that activate together connect jointly. More rigorously, two neuronal groups A and B are connected if the action potential, or discharge, needed in one group to react with the other is weaker than for other neuronal groups. It operates as though preferential pathways form between brain cells through experience, resistance to information flow reducing through habituation—like a path's grass thinning from foot traffic.

Of course, brain evolution isn't purely chance-driven. Neuronal excitations, the second ingredient of Edelman's Darwinian theory, respond to environmental constraints. Connexion between neurons, or neuronal groups, is encouraged if physiological responses resulting from this activity correlation appropriately support organism needs and survival in nature. A typical example is memory function in aplysia (a giant marine snail), studied by Kandel[11]. When this animal receives a shock to its body, like the tail, it

[11] Eric R. Kandel, *The Molecular Biology of Memory: A Dialogue Between Genes and*

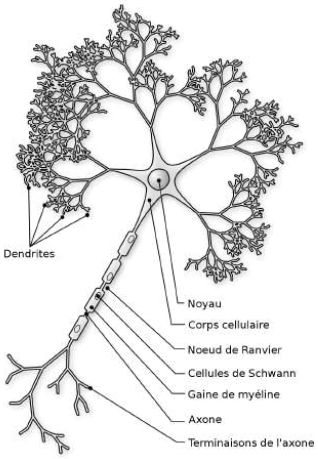


Figure 14.1: Schematically, a neuron primarily consists of a nucleus, dendrites, and a long axon. Dendrites receive afferent signals. If their summation activates the neuron, an electrical potential communicates along the axon via Schwann cells to other neurons. Receiving a neural impulse can be facilitated or inhibited by neurotransmitters (image source: Wikipedia, French article on neurons).

recognises the stimulus' unpleasantness and learns to enhance its defensive reflex to various subsequent stimuli, manifest in muscle contraction, withdrawal posture, greater reactivity, etc. "A single shock produces a memory lasting only minutes; this short-term memory does not require new protein synthesis. However, four or five spaced shocks on the foot produce a memory lasting several days[12]." This results from an indirect connexion between sensory and motor neurons, mediated by a small group of excitatory and inhibitory interneurons.

Remaining is an explanation of the mechanism ensuring proliferation of neurons, or neuronal connexions, selected upon environmental contact. In the aplysia learning framework, Kandel's work illustrated that trauma memory is possible via a mechanism involving specific protein (CREB-1) creation in nerve cells, mediated by messenger RNA. Protein creation distinguishes short-term memory from long-term memory. Edelman points to another phenomenon: re-entry. Re-entry involves ongoing signal passage from one brain region (or map) to another, and their return, through massively parallel fibres (axons) known for their ubiquity in advanced brains. This provides the explanation for observing simultaneous activation of different brain parts in the absence of a central administrator or any organ tasked with encoding top-down brain function. To Edelman, it functions as

Synapses, M/S: médecine sciences, 19(5), 625–633.
[12] *Ibidem*, p.627

the central argument refuting cerebral function assimilation to a computer's. To illustrate re-entry, the Nobel laureate offers an imaginative analogy:

> *To imagine how re-entry works, consider a hypothetical string quartet composed of talented musicians. Each plays their air at a different rhythm. Now connect the bodies of all musicians with tiny wires (numerous on all body parts). Whenever each musician moves, they unconsciously send movement waves to others. After a short while, the rhythm and somewhat the melodies become more consistent. This dynamic proceeds, producing new coherent output. It's akin to what happens in improvised jazz, sans wires, of course*[13] *!*

Re-entry, according to Edelman, accounts for forming coherent ideas absent a regulatory organ. One recalls that, for Descartes, the pineal gland played a central role synthesising duplicated sensory information (the two eyes, ears, nostrils...) into a homogeneous, clear, and distinct idea of perceived objects. However, under Edelman's theoretical model, covering for the pineal gland or any central appendage has become redundant. Conversely, unified vision perception arises from simultaneous coordination of brain areas related to stimulus orientation (area V1), colour (V4), and movement (area V5).

Note that re-entry has found considerable traction in contemporary neuroscientific theories, forming not only the core of Giulio Tononi's theory—which will be explored in the next chapter—but also in the consciousness models of Dehaene, Changeux, or Naccache.

14.4 Edelmanian Theory of Consciousness

How might the Darwinian neural theory just presented support a scientific or philosophical theory of consciousness? If applicable, what does such a theory imply regarding faculties associated with consciousness, such as reason, imagination, intuition, or creativity? Before tackling these central issues, Edelman clarifies that phenomenological consciousness, defined as the momentary experience of qualitative and subjective life, should not be confused with self-consciousness or reflexive consciousness, which involves being aware of oneself as a thinking subject. This parallels the distinction between "I think, I am" as an irreducible fact of consciousness and the very act of thinking: "I think, therefore I am."

In a work co-authored with Tononi[14], Edelman proposes that consciousness arose when the thalamocortical system significantly expanded and the

[13] Gerald M. Edelman, *Second Nature: Brain Science and Human Knowledge*, Yale University Press, p. 44

[14] Gerald H. Edelman, G. Tononi, *How Matter Becomes Conscious*, Odile Jacob, 1st May 2000.

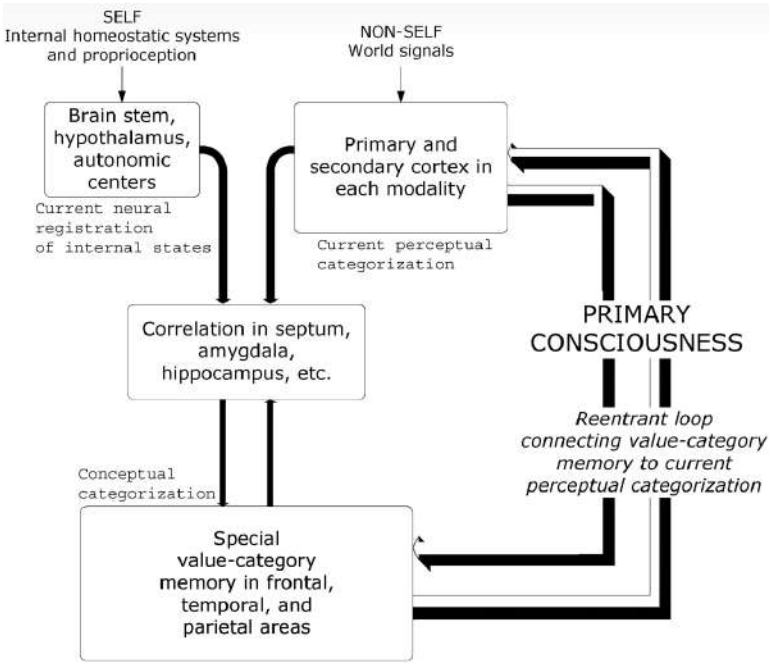


Figure 14.2: Example of re-entry system proposed by Gerald Edelman (source: Gerald M. Edelman, *Naturalizing consciousness: A theoretical framework*, PNAS, April 2003).

number of specific thalamic nuclei increased with the expansion of the cerebral cortex. These conditions indeed promote reentrant activity between cortical areas and the thalamus. It is posited that this dates back to the transition between reptiles and birds, then mammals, nearly two hundred and fifty million years ago. While the anatomical argument is justified by concepts presented in the preceding chapter, it is challenging to see how this assertion holds scientific value. Moreover, this raises questions about the experiences of dinosaurs or animals from the Cambrian explosion. Perhaps it is necessary to distinguish sentience from true consciousness? Perhaps Edelman's theory, while explaining human consciousness, does not preclude that consciousness experiences, though different from our own, could arise from other physiological mechanisms than those specific to our brains?

Edelman is more optimistic than his counterpart Julian Jaynes, who traced consciousness back to a period closely following the emergence of language and metaphors[15], even seeking evidence in Homer of a mental life very different from what we know today.

According to Edelman, animals with a brain capable of strong reentrance can categorise sensory information finely and associate synaptic values corresponding to each of these categories with certain memory areas. When a scene is perceived at a given moment, categories are extracted by the thalamocortical brain, while memory mediation is facilitated by more anterior cortical regions, such as the frontal and parietal cortex. In this way, it is capable of structuring incidence information to form unified percepts: instead of perceiving the world as a rhapsody of sensory data crashing into the human mind like a tide against a breakwater, the contours of determined objects come to life amidst the waves. It would be intriguing to consider what the sensory world would resemble without these percepts? Would it be a calm sea, homogenous at every point in space, a sort of Parmenidean being without movement, without asperities? Or rather a storm worthy of the primordial chaos of great mythologies? This question will be addressed in the chapter...

This described consciousness could be qualified as primary, as it focuses so to speak only on present states. From a neuroscientific perspective, Edelman suggests that self-consciousness appears by adding the brain's ability to reflect on its own states.

[15] In his book *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Julian Jaynes traces the emergence of self-consciousness to an era around the second millennium BC. He stipulates that before this date, a bicameral mind prevailed, inspired by the two cerebral hemispheres: one part of this mind consisted of voices, and the other part listened to them. The notion of the self as a unitary and autonomous entity was fragmented between these two instances.

14.5 Metaphysical Chiasms and Epistemic Principles

Although Edelman denies all forms of dualism, or perhaps because of this, his theory, to achieve full meaning, indirectly relies on common sense psychology. When he writes, to describe the phenomenon of memory, that "the brain talks to itself[16]", he weaves an implicit correspondence between a physiological phenomenon, involving many Darwinian biological mechanisms, and an experience lived by a human subject. Generally, Edelman thus provides a scholarly description of brain function as an organism, but he does not truly address the hard question of consciousness: namely, why is it that one feels something rather than nothing? Why is it that from the synergy of all these biochemical mechanisms a conscious phenomenon, so to speak, arises?

Edelman should not be labelled as philosophically naive—an epithet particularly difficult to attribute as it constitutes, for the contemporary academic, a kind of walking nightmare. Although he challenges dualism and attempts to construct a theory of consciousness without explicitly resorting to introspection, Edelman concedes that for the theory of consciousness to gain significance, it must be based on dualism. What he simply denies is that the mind as such can have a causal action on matter. Or put differently, he does not wish to endow abstract entities, such as will, self, or memory, with any real significance beyond the ideas humans form of them. In this, he certainly opposes Descartes' pineal gland, moved by the soul's sole force, but he also aligns with the position I seek to defend, within which the modalities of body and mind do not have a direct action on each other: they relate to one fundamental reality.

> *To comprehend that it is the reentrant and selectionist activity of the neuron groups forming part of the nucleus that leads to phenomenal consciousness, it is necessary to evoke dualism. Even though consciousness is a process without causal power in itself, it is nevertheless driven by the complex activity and causal power of the neuron groups constituting the reentrant nucleus. Furthermore, from the earliest stages of development, the signals from the body to the brain and from the latter to itself form the basis for the emergence of a self*[17].

Edelman speaks of causal power, suggesting that for him, brain activities can cause mental states, such as perception or self-consciousness. However, he clarifies (a) that these mental states, or qualia, are not things of the brain, but indeed physiological processes related to neuronal information

[16] Edelman, *op. cit.*, p. 47

[17] Edelman, *Second Nature: Brain Science and Knowledge*, Odile Jacob, p. 53.

processing and (b) that the notion of cause is not entirely accurate, and we should rather speak of a support for the emergence of the phenomenon: "The hypotheses that it is the action of the thalamocortical nucleus that causes consciousness encounter a difficulty. Since causes precede effects, these hypotheses imply the existence of a temporal gap between incommensurate processes. Conversely, neural action in the nucleus induces consciousness, much like the spectrum of blood haemoglobin is induced by the mechanical and quantitative structure of this molecule[18]."

In his book *Wider than the Sky: A New General Theory of the Brain*, Edelman even suggests a psychophysical parallelism, with on one side the mental processes as we experience them, filled with memories, experiences, representations—or, in a word, qualia—and on the other side the cerebral processes composed of neurons, myelin, axons, and other cerebellous substances[19]. Although there is no direct causality from mind to matter, or from matter to mind, Edelman argues that the identifiable causes between neurons correspond, analogously so to speak, with the causes between mental phenomena. This assertion is very close to the structural monism I advocate in this thesis, but it presents two major differences.

On one hand, it continues to view mind and matter as full-fledged substances (the mental and the material), at the risk of contradicting Edelman's idea that consciousness is not a thing, but a process. On the other hand, the mind is conceived in somewhat naive terms. It is assumed that the Self, the core of consciousness, is the cause of the production of thoughts and sensations. In either case, there is a reification of consciousness. It is viewed as an objective entity acting in the world. Contrary to this approach, I defend the thesis according to which consciousness is a phenomenon that is lived but cannot be represented as an object within the mental world. Far from being at the core of mental life, consciousness is a horizon, a boundary that cannot be directly grasped, but whose concept is nonetheless necessary and meaningful.

The idea of parallelism is enticing, yet ideally, a more in-depth analogy should be drawn between the experience of memory, creativity, imagination, as we live them, and the principles underpinning neuronal function, to demonstrate that, in the end, it is the same processual structure. This could be accomplished by developing, in parallel with brain theory, a theory of lived consciousness, showing that this theory aligns with the same epistemological principles as the theory of neuronal group selection. This parallelism would not presuppose two distinct substances with their causal powers, but the existence of a single psychophysical reality that can be viewed from two distinct perspectives.

Such an approach will be considered in the third part of this thesis. For

[18] *Ibidem*, p. 57.

[19] Refer to the chapter *Phenomenal Transformation* in the aforementioned work.

now, I would like to summarise here the overarching theoretical guidelines of Edelman's theory[20]:

- **Enactivism:** Edelman views consciousness as the product of three ingredients: the presence of a brain, the fact that this brain is enveloped in a body, and the interaction of this body with the environment. In this regard, he opposes a computational vision that likens the operation of the mind to that of a computer. There is no centralised clock in the brain organ, allowing for a certain number of logical operations per unit of time. However, Edelman does not deny the crucial role of the brain in the phenomenon of consciousness.

- **Evolutionism:** For the same reason that the brain is not a computer, its functioning is not preprogrammed by the genetic material of the individual. Neural networks, quite to the contrary, are selected in interaction with the environment through a Darwinian process. This is enabled by three factors: (1) pre-existing neuronal diversity; (2) the strengthening of synaptic connexions thanks to Hebb's rule, which states that neurons that fire together, wire together; (3) the presence of reentrant activity, ensuring strong integration of neurons in a broader environment. "Reentry," says Edelman, "is the ongoing recursive exchange of parallel signals between brain areas that serves to coordinate the activity of different zones in space and time. Unlike feedback, reentry is not a sequential transmission of an error signal in a simple loop. It simultaneously involves many reciprocal parallel pathways and lacks a prescribed error function attached to it[21]".

- **Processual Emergentism:** Edelman borrows from William James the idea that consciousness should not be seen purely as a thing, but rather as a process. Unlike logical or psychological atomism, it is not a matter of reducing consciousness to as many cognitive or psychological abilities (reflex, sensation, vision, language, memory...) and then associating these abilities with parts of the brain (sensorimotor cortex, Broca's area, Wernicke's area, hippocampus...). Consciousness is instead a process within a complex system. "A complex system," Edelman explains, "is a system that consists of various small parts, each of which can be functionally isolated. As these heterogeneous

[20] For a more detailed discussion of Edelman's epistemology, I refer readers to the article by Edmond Ortigues titled *Edelman and Human Consciousness*, Éditions de Minuit, in Critique 2005-12 (Num. 703), 2005-12 (Num. 703), pages 917 to 926.

[21] Gerald M. Edelman, *Wider than the Sky: A New General Theory of the Brain*, Odile Jacob, Paris, p. 57.

parts interact in varied combinations, there is a tendency to give rise to more integrated properties of the system[22].».

- **Emergent Causality:** Edelman's stance on causality is tentative, but it is clear that he opposes the idea that matter acts on the mind, or conversely that the mind acts on matter. He seems to view neuronal substrates as a support for the emergence of mental phenomena. In this regard, he seems to align with an emergentist theory reminiscent of the idea of supervenience advocated by Jaegwon Kim[23]. In certain instances, he prefers to speak of psychophysical parallelism, which more closely resembles a Leibnizian conception.

[22] *Ibidem*, p. 87

[23] Notably see his article *Making Sense of Emergence*, as well as his book *Supervenience and the Mind*.

Key Takeaways from this Chapter

Gerald Edelman is renowned for applying his theory on the selection of immunoglobulins, which earned him the Nobel Prize in 1972, to the study of consciousness.

The theory on immunoglobulin selection posited that antibodies are originally present in the blood in varied forms. In the presence of a pathogen, those that have the form required to counter the threat will quickly multiply, resulting in a *de facto* increased resistance of the organism to subsequent infections.

Maintaining the Darwinian spirit of this mechanism, Edelman makes three hypotheses about the human brain:

1. The population of neurons initially presents diverse shapes and connexions.
2. In the presence of environmental stimulation, certain neurons will be excited simultaneously, others not. Neurons that excite simultaneously will statistically bond (Hebb's law).
3. Linked neuron groups will exhibit significant "reentrant" activity: signals entering a cerebral system will pass into other systems before being reinjected into the initial system, recursively so.

According to Edelman, consciousness should not be conceived as a product of the neurons themselves, regarded as physical objects, but as the process that these neurons support. It emerges from circular information processing.

Edelman thus advocates a non-reductionist view of consciousness, tinged with enactivism. However, he assumes an implicit causation between lived conscious phenomena and brain activities. In this respect, it is permissible to suspect, within his thought, the presence of an externalist metaphysical chiasm.

Chapter 15

The Integrated Information Theory

> Consciousness is not an object;
> it is a process.
>
> Giulio Tononi

The history of thought presents itself as a grand, flat canvas, which ideas of varying depths and meanings come to warp in turn. Each contribution gently deflects the field of knowledge, at times creating fertile grounds, at other times inescapable chasms. In this regard, the theory of Gerald Edelman is no exception. On one hand, it opens a new field of inquiry, where consciousness is conceived as an emergent evolutionary phenomenon, resting on the high integration of neural networks, rather than merely a cognitive property, localisable to a brain region much like language production was once associated with Broca's area, language comprehension with Wernicke's area, or memory with the hippocampus.

Yet, this promising terrain, akin more to a fallow field than a harvest-ready land, is sorely lacking in landmarks. Even though Edelman emphasises the role of the thalamocortical system, serving as a relay for sensorimotor information between the cortex and the thalamus[1], he offers no definitive criteria to identify the higher-order structures involved in the phenomenon of consciousness. Are brain-injured patients, with only a small portion of their cortex remaining intact, conscious? Are newborns conscious? Are animals displaying complex behaviours, yet having brains very different from humans, conscious? Can intelligent machines be conscious[2]?

[1] In his technical work *The Remembered Present: A Biological Theory of Consciousness*, Gerald Edelman locates the origin of the biological self in the limbic system and brainstem, which endows the organism with goals and values, while attributing to the thalamocortical system the capacity to perform continuous comparisons between the self and the non-self.

[2] These questions are examples posed by Tononi and colleagues in their opinion arti-

Edelman pondered early on the existence of a physical substrate for consciousness. In 1999, he published an article with Giulio Tononi titled *Consciousness and Complexity*, which laid the foundation for what would later become the integrated information theory. They took a dual stance: to move beyond the holistic position of William James, who believed the entire brain was responsible for consciousness[3], and also beyond the localisationist position of some neuroscientists, who attribute to specific neurons or neuron groups the power to create conscious phenomena. This latter approach, in their words borrowing from Gilbert Ryle, represents a genuine "category mistake"[4], ascribing a genuinely mental and sensory power to material units. To avoid this pitfall, what middle path remains to be taken?

The authors suggest focusing not on particular groups of neurons, nor even brain regions, but on types of neuronal functioning. Assuming consciousness is primarily an information-processing process, not an organ or object among others, their strategy is to isolate the neuronal mechanism most likely to embrace the properties of this process, and then, in a second phase, to seek where in the brain this mechanism can be observed.

This leads to a significant nuance with the purely localisationist approach: first, it is not necessary for the same brain regions to implement the mechanisms of consciousness at all times. Conversely, it is entirely possible that, during sleep, different parts of the brain take over from those of wakefulness to manage conscious mechanisms. Furthermore, variations may exist depending on the types of tasks performed (visual, auditory, memory...), the individuals, or the presence of lesions. Finally, it is theoretically admissible that these processes are implemented by substrates other than those the human brain has familiarised us with—consider insect brains or silicon machines.

To identify these neuronal processes, Edelman and Tononi adopt a reversed methodology. Instead of starting from the brain and gropingly searching for the presence of conscious activity, they take phenomenological experience as a starting point, hoping to discover criteria that will only then serve as indices for identifying appropriate neuronal processes. The initial research question thus becomes: what does it mean to have a conscious experience?

Edelman and Tononi respond by isolating two key components in neuronal processes: *integration* and *differentiation*. Integration refers to the fact that mental contents are presented to consciousness in a unified way, that is, as an indivisible totality. It is a scene, certainly mobilising different internal and external senses, yet within which everything is presented as one single block. In this respect, the individual is compelled, as it were, to view

cle titled *Integrated Information Theory: From Consciousness to its Physical Substrate*, Nature Review Neurosciences, May 2016.

[3] James Williams, *The Principles of Psychology*, Holt, New York, 1890.

[4] Ryle G., *The Concept of Mind*, Hutchinson, London, 1949.

the world from a private, intimate perspective.

Differentiation, on the other hand, consists in that, although every act of consciousness presents as a unified whole, the mind is capable of decomposing this whole into distinct elements, proceeding in this way through multiple comparisons. The time interval required to decompose a scene is extremely short. For visual discrimination alone, a fraction of a second suffices to distinguish, within an image, a plethora of distinct objects. Thus, there is a process present that differentiates information, in such a way as to isolate concepts or patterns within the whole.

Once the phenomenal properties have been identified, they remain to be formulated in a quantitative language, enabling the determination of whether, yes or no, a neuronal process can be qualified as conscious. To this end, Tononi and Edelman resort to entropy, a measure developed by Shannon to gauge the amount of information between several statistical variables[5].

More specifically, they estimate that each neuron behaves like an *a priori* random variable, which can take the value 1 (excitation) or 0 (rest) and exerts a probabilistic influence on the activity or rest of other neurons. The question of whether a set of neurons forms a strongly integrated unit thus reduces to asking whether it presents strong statistical independence from the rest of the neuronal system. This can be determined mathematically through an entropic function called mutual information, which measures the loss of uncertainty acquired on a set of random variables due to the information held about another set of random variables[6].

Building on these initial works, carried out in concert with Edelman, Tononi continues to refine the theoretical approach through several successive versions of the integrated information theory 1.0, 2.0, 3.0, and 4.0. He gradually provides his theory with an epistemological foundation, aided by four ingredients: (1) firstly, he defines a set of methodological principles based on a form of operational reductionism; (2) he gives axioms to describe as simply as possible what constitutes an authentic experience of consciousness, and (3) he matches these axioms with physicalist postulates, each endowed with a quantitative measurement tool. His core idea is to establish an analogy between the phenomenological functioning of conscious experi-

[5]For more details on the technical tools, refer to appendix 6.

[6]The problem of isolating the statistically independent components of an information set dates back to the *cocktail party problem*: when a group of individuals is gathered in one space and converse in small groups, how does the brain isolate the conversation's pertinent information from the ambient noise? The neurological explanation is that humans are endowed with two receptors (the two ears). It is possible to cross the sound information from each ear to isolate a conversation from the ambient noise. This conversation is indeed statistically independent of the surrounding noise. The mathematical measures used to model this kind of problem sometimes rely on independent component analysis (based on mutual information), principal component analysis, or simply the Kullback-Leibler divergence.

ence and the hidden physical mechanisms behind certain neuronal processes. This leads him to discuss certain consequences of his theory concerning ontology[7], free will[8], and information[9].

In this chapter, I will provide a substitute for Tononi's theory in its latest version (version 4.0, released in December 2022), after which I shall undertake a philosophical perspective.

15.1 Strategy of Structural Equivalence

Tononi seeks to establish a theory of consciousness that, like Gerald Edelman, takes phenomenological experience seriously, understood as the irreducible act of feeling something rather than nothing. In the introduction to his article titled *Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms*, he writes: "To be conscious—to have an experience—means 'it feels like something to be conscious': it feels like contemplating the sky, hearing the roar of the ocean, dreaming about a friend's face, imagining the flow of a melody, considering a choice, or reflecting on one's own conscious experience."

Thus, Tononi believes that in the absence of an explanatory link, merely describing the neurological properties of the brain is insufficient to fully understand the phenomenon of consciousness. Yet, he does not wish to establish a Cartesian dualism that would oppose, on one hand, the modality of consciousness and, on the other, that of the body. In keeping with the scientific spirit, he seeks to account for conscious experience in physicalist terms, but in such a way that a light or a sense can be given to the experience of consciousness.

To achieve this, he begins by defining a set of axioms describing the fundamental properties of phenomenological experience, followed by a set of postulates regarding the physiological functioning of the mind itself, envisioned as a particular way of processing information. The aim then becomes to demonstrate that the axioms of phenomenological experience virtually coincide with physicalist postulates; the former can be explained concerning the latter and vice versa. This approach, which might be termed *structural analogy*—and which Tononi calls an *explanatory identity*—rests on five epistemological principles:

1. **Inference to the Best Explanation:** The postulates of the mind are hypotheses chosen to most effectively account not only for known neurological phenomena but also for the axioms of phenomenological

[7] Tononi G., *On being*; forthcoming.

[8] Tononi G. et al., *Only what exists can cause: An intrinsic view of free will*, arXiv, June 2022.

[9] Tononi G. et al., *System Integrated Information*, Georg Northoff, 11 February 2023.

experience. Ideally, the resulting theory should embrace a large number of facts (breadth), in a unified manner (synthesis), explaining the facts precisely (specificity), without internal contradiction (systemic consistency), be subject to experimental predictions (scientific validity), while remaining simpler than rival theories (simplicity)[10].

2. **Realism:** It is hypothesised that something exists and persists outside conscious experience. The experience of the external world serves as a condition for the elaboration of physicalist postulates. Tononi further holds that the hypothesis of realism, unlike solipsism (which denies external reality), is the only one capable of giving rise to explanations and predictions about the world. It is possible that this assertion conceals a certain metaphysical naivety. One could imagine an idealism in which reality would be entirely contained within the mind, though this mind itself would be governed by stable and universal laws. However, the hypothesis of realism is not surprising within a scientific approach, yet I wonder to what extent it proves indispensable to Tononi's theory?

3. **Operational Physicalism:** To determine what exists outside phenomenological experience, Tononi and his colleagues believe it is necessary to resort to an operational criterion: Does the existence of the physical entity lead to a reliable and predictive explanation of phenomenological experiences? For instance, one could imagine that the existence of a tree in the world is validated if this idea accounts for the repeated experiences of the table in our consciousness (whenever I have looked this way, I have experienced a tree), and it also allows for predictions (when I look this way again, I will see a tree). In the context of a systematic theory, however, the objects handled are not tables but elementary units: electrons, neutrons, hadrons, etc. But since we are here concerned with explaining what consciousness is, these units are seen as irreducible elements of neuronal information.

4. **Operational Reductionism (atomism):** According to Tononi

[10] One must be cautious that inference to the best explanation (IBE) is often confused with abduction. The former consists of holding true the hypothesis that best explains the hypotheses, whereas the latter, according to Peirce, postulates making the hypothesis of unobservable causes such as the presence of invisible particles or the existence of a bygone historical event. While the first IBE limits itself to observations and is thus more secure, abduction relies on imagination. It is less certain, but can prove more fruitful (fertility). For further details regarding this distinction, refer to Frédéric Roudaut, *How Hypotheses are Invented: Peirce and the Theory of Abduction*, in Cahiers philosophiques 2017-3 (Num 150), pages 45 to 65.

and his colleagues, in principle, there is no need to postulate anything other than the existence of elementary units and causal relationships governing them. From a mathematical standpoint, the ultimate task of physics is thus to define a probabilistic transition matrix $T_{\mathcal{U}}$ stipulating, for each state u of the universe at a given moment (state u consisting of a certain arrangement of all elementary units), the probability of observing state u' in the subsequent moment. Furthermore, the observer must be able to manipulate each elementary unit individually to determine to what extent modifying one of them impacts the other units of the system. In this sense, to conclude, we can say that Tononi's epistemology consists of flattening physical phenomena. The concepts of universal law or physical property are rendered obsolete. There remain only the entities whose manipulation induces observable differences in predicting the system's evolution. Note, however, that this reductionism is called *operational*, not *ontological*, because these elementary units are not necessarily viewed as realities *of the world*, but rather as components necessary for theorisation. Tononi indeed argues that only consciousness itself and the corresponding information processes truly exist, not individual neurons (which exist for consciousness, that is, extrinsically).

5. **Intrinsic Perspective:** To account for conscious experience in physicalist terms, the existence of something must be assessed from the intimate perspective of the subject, not from the viewpoint of an external observer. In the terms I introduced, it means considering a perspective in I rather than He. This is conceived by Tononi as a set of mathematical properties presented by the physical system. This implies, notably, that the information available to the mind must depend on its own state, not on the average of states observed by an external observer. Ontologically, Tononi even goes so far as to say that only complex systems with an absolute maximum density of information truly exist. These systems correspond to conscious processes, not to the gravitation of stars in the gaping night space or to the climatic depressions vainly agitating the atmosphere.

Despite being methodically introduced by Tononi and his colleagues at the start of their article, these principles lack any philosophical justification. They are working hypotheses to which the reader is asked to adhere, at least provisionally, before embarking on a reading of the system in its entirety. Besides the lack of justification, there is also a number of blind spots. For example, I regret the absence of a formulation of the principle that consciousness results from information processing. This idea seems crucial in Tononi's

thought, and although adopted by many scientists, including Edelman, it is debated by the scientific community. Among the advocates of this idea, we find authors such as Baars[11], Aleksander and Gamez[12], Earl[13], Jonkisz[14], Fingelkurts and Fingelkurts[15] or more recently, Giorgio Marchetti[16].

Furthermore, it might be instructive to link this idea with an axiom of phenomenological experience: what, in being conscious, as we experience it daily, can be analogised to an information processing mechanism? I will advocate in chapter ... that lived consciousness relies on a principle of differentiation. Consciousness is not populated with entirely stable objects, given once for all in their entirety, but proceeds by comparison. In a ceaseless play of approximations, differentiations, analogies, and symmetries, it carves categories manipulable from the very fabric of sensory experience: shapes, colours, patterns, concepts, all the distinctions language has familiarised us with, and that we take as given facts, result from an incessant movement of perception. I will call this principle the *principle of differentiation*. But for now, I would like to delve deeper into Tononi's theory and examine his axioms of phenomenological experience.

15.2 Axioms of Phenomenological Experience

Tononi's approach begins by defining, through five axioms, the essential properties and structure of conscious experience. These axioms, in his view, are propositions held to be true in themselves, which cannot be inferred from other propositions and thus constitute starting points for logical inference. Tononi provides no further justifications for these axioms, but he seems to regard them as the product of a certain intuitive evidence. "Phenomenological existence is immediate," he writes, "in the sense that my experience is simply present, directly far more than indirectly: I do not need to infer this existence from something else[17]"

[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

[17]Tononi et al., *Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms*, p.3.

In this respect, his vision of axiomatic might be likened to that of Frege when he writes: "I call axioms propositions that are true but cannot be demonstrated because the knowledge we have of them depends on a source of knowledge totally different from the logical source, which can be called spatial intuition. From the truth of the axioms, it follows that they do not contradict themselves[18]." However, the type of intuition at play in the case of consciousness is not spatial intuition but a properly existential intuition.

I will list the five axioms in question and comment on them to show that the intuitive evidence underpinning them may not be as undeniable as it might appear at first glance[19]

- **Axiom 1 (Existence):** Consciousness exists. Every experience is real – indeed, the fact that my experience here and now exists (it is real) is the only fact of which I can be immediately and absolutely certain. Moreover, my experience exists from its own intrinsic perspective, independent of external observers (it is intrinsically real or actual).

- **Axiom 2 (Information):** Consciousness is specific: every experience is what it is – being composed of a specific set of specific phenomenal distinctions – thus differing from other possible experiences (differentiation). For example, an experience may include phenomenal distinctions specifying numerous spatial locations, several positive concepts like a room (as opposed to no room), a bed (as opposed to no bed), a book (as opposed to no book), a blue colour (as opposed to no blue), higher-order "bindings" of first-order distinctions, like a blue book (as opposed to no blue book), as well as numerous negative concepts, like no bird (as opposed to a bird), no bike (as opposed to a bike), no bush (as opposed to a bush), etc. Similarly, an experience of pure darkness and silence is what it is: it has the specific quality that it possesses (no room, no bed, no book, no blue, nor any other object, colour, sound, thought, etc.). And it is thus different from a large number of alternative experiences I might have lived but am not actually living.

- **Axiom 3 (Integration):** Consciousness is unified: every experience is irreducible to non-interdependent and disjoint subsets of phenome-

[18] Cf. letter XV/3 of 27.XII.1899 in Gottlob Frege Briefwechsel mit D. Hilbert, E. Husserl, B. Russell, and selected individual letters from Frege, edited by G. Gabriel, Fr. Kambartel, Chr. Thiel, Hamburg, Felix Meiner, 1980, p. 9 (modified French translation in Logic. . . , p. 223).

[19] The formulation of the axioms is taken from scholarpedia and not from the aforementioned article by Tononi, which presents them somewhat too succinctly. See here: ...

nal distinctions. Thus, I experience an entire visual scene, not the left side of the visual field independently of the right side (and vice versa). For example, the experience of seeing the word "BECAUSE" written in the middle of a blank page is irreducible to an experience of seeing "CAUSE" on the right plus an experience of seeing "BE" on the left. Similarly, seeing a blue book is irreducible to seeing a book without blue, plus blue without the book.

- **Axiom 4 (Exclusion):** Consciousness is defined, in terms of content and spatiotemporal grain: every experience possesses the set of phenomenal distinctions it possesses, no less (a subset) and no more (a superset), and it flows at the speed it flows, neither faster nor slower. For instance, the experience I live is of seeing a body on a bed in a room, a bookshelf with books, one being a blue book, but I do not live an experience with less content – say, devoid of the phenomenal distinction blue/non-blue, or coloured/non-coloured; or with more content – say, endowed with the additional phenomenal distinction hypertension/hypotension. Moreover, my experience unfolds at a particular speed – each experience lasting, say, about a hundred milliseconds – but I do not live an experience that lasts only a few milliseconds, nor rather minutes or hours.

- **Axiom 5 (Composition):** Consciousness is structured: every experience is composed of multiple elementary or higher-order phenomenal distinctions. For example, during an experience, I can distinguish a book, a blue colour, a blue book, the left side, a blue book on the left, and so forth.

As an example, Tononi and his colleagues suggest picturing a sleeper who, suddenly awakened from dreamless sleep, experiences the white wall of his room, his spring-laden bed, and his semi-lethargic body. This experience exists immediately and irrefutably, yet it exists for him (Axiom 1) specifically: the wall is a wall and it is white (Axiom 2). Moreover, this experience is unitary, in the precise sense that the room's left side is not experienced, within this experience, separately from the right side (Axiom 3). The experience is also bounded in space and time: the sleeper does not have access to what is behind my head, but I cannot subtract what falls under my eyes (Axiom 4). Finally, the phenomenological experience, according to Tononi, is structured by a set of distinctions (the wall, the bed, the body) and relations (the body is on the bed, the bed in the room), engaging as many concepts (Axiom 5). In this last illustration lies a conception of consciousness reminiscent of Edelman, who said of it: "Consciousness is

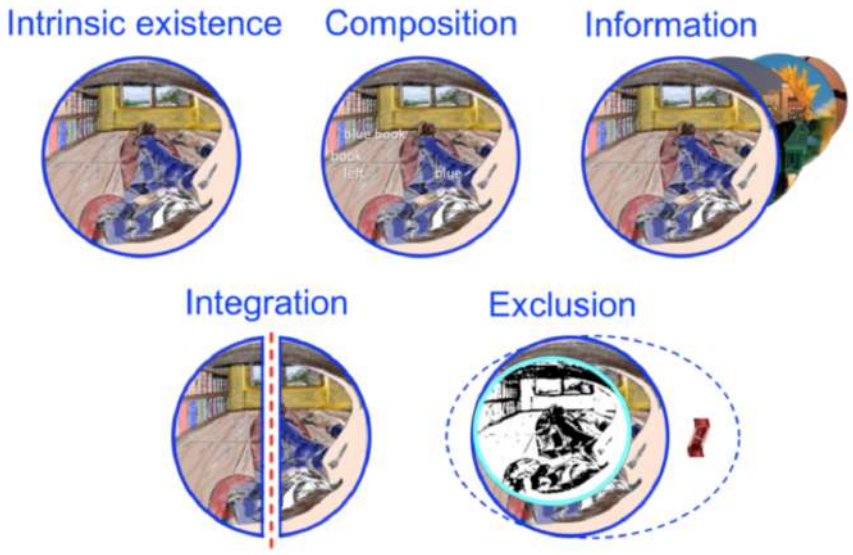


Figure 15.1: This figure illustrates the five axioms of phenomenological experience from the view of a scene.

the aptitude to construct a complex scene and to perform discriminations among the components of this scene[20]."

Despite this illustration, numerous questions remain unanswered. Many of them relate to the exact definition of words. The interpretation of terms such as "existence," the "whole," or "structure," laden with philosophical weight, leads to uncertainties. Axiom 3, for example, stipulates that experience must be considered as an irreducible whole, within which the lived experience cannot be reduced to a set of separate experiences. Yet, Axiom 5, on the contrary, states that experience is "structured" by a set of "distinctions." Isn't making distinctions precisely separating parts of the phenomenological experience (the wall, the bed, etc...) from the whole of perception (the room in its entirety)? How can the lived experience be both one and multiple? And where does this structure originate?

To understand this objection, I wish to resort to a thought experiment. Suppose a mature man who, following a spectacular head trauma, retained all his cognitive faculties... except memory. This man no longer knows what he is, what he is doing there, or how to name the objects around him. If you present yourself before him, he will barely recognise you as one of his kind—only to forget you instantly.

This forsaken soul, doomed to eternal oblivion, is nonetheless aware of the world around him. But as his attention leaps from one idea to another, he forgets the previous idea with each movement, so that life, for him, is

[20] Gerald M. Edelman, *Wider than the Sky: A New General Theory of the Brain*, Paris, Odile Jacob, 2004, p.76.

reduced to a unique, ever-renewed instant. At times, he is contemplation of the landscape; at times, he is pain; he is sound, smell, colour, tree, mountain, faces. He entirely immerses himself in the lived world, retaining no awareness of his own existence. Every moment of life is the birth and death of a new world. It is continuous creation, without continuity between different creations.

The experience of this brain-injured patient is indeed far removed from what we experience daily, and perhaps it would be impossible to form a coherent idea of it. Can we say that this man is conscious of the world otherwise than as an undifferentiated whole? When we inspect, within the room, the presence of a chair, it seems we detach this isolated perception from the whole to which it relates. To do so, our attention, previously directed at the room as a whole, must shift towards this singular chair; so the distinction between the room and its parts induces a temporality.

Initially, our mind is pure experience of the room; in the next moment, it is contemplation of the chair. The essential question of any phenomenology of consciousness then becomes how isolated moments of consciousness relate to each other coherently, rather than being experienced as disconnected phenomena?

This question is none other than the question of memory: each moment of conscious experience retains within it, or refers indirectly to, previous moments, weaving as many structural links between successive ideas, perceptions, and emotions. This very faculty differentiates us, healthy human beings, from the case of the brain-injured patient I mentioned earlier.

From this, we conclude the following: if the experience of consciousness were truly unitary, it would be such that, within it, distinctions between objects, concepts, and affects typically manipulated by the human mind would disappear. This myriad of categories we are familiar with, carving reality with their sharp blades, would fade surreptitiously, leaving nothing but an undifferentiated apprehension of the present moment.

Indeed, it may be impossible to grasp perfectly what such an experience of *pure contemplation* would be, just as we cannot fathom what life would be for a being devoid of memory. Yet, it is conceivable that borderline cases of consciousness could fall under scientific observation. I am particularly thinking of the deconstructions observed in certain psychotic episodes. In his work titled *Dementia Praecox or the Group of Schizophrenias*, Breuer argues that the primary symptom of schizophrenia is the alteration of connexions ordinarily linking associations of ideas:

> "The connexions 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).

In a schizophrenic crisis, the successive perception of the room, the chair in its centre, the white wall or one's own body does not lead the subject to systematically construct a coherent environment, but divides the wounded consciousness into all these indivisible experiences, where the distinction between observer and observed object, between observed object and other objects, emotions, or relations of mental life, dangerously fades. Connexions exist indeed, but they are delusional. They are, so to speak, the ultimate twitch of a consciousness unable to give structure to its unfolding. The subject may feel as if he is the chair he is observing, just as he might attribute paranoid affective connotations to certain perceptions. In this respect, one might suggest that schizophrenia essentially presents as a disease of concepts, specifying, however, that these concepts should not be confined to formal reasoning, but concern the entire mental life of the subject: the logical, perceptual, affective, social, and relational spheres. The very fact that a paranoid patient continues, in the absence of any capacity to do so, to weave coherent relationships among things, shows how fundamentally consciousness is a capacity to connect and differentiate experiences.

Tononi's Axioms 3 and 5 do not, therefore, seem erroneous, but they are content to describe, as settled facts, some key properties of consciousness without delving into the dynamics linking perception, temporality, memory, and totality. Thus, it becomes impossible to understand, in non-paradoxical terms, how consciousness can appear, *at the same time*, as a totality and a fragmentation. To put it succinctly, all indicates that Tononi and his colleagues have not undertaken an in-depth reflection on the intimate experience of consciousness. They seem to consider that the mental life of the individual, unlike other objects of scientific interest, does not require revelation through specific experiences, but is displayed almost directly via the magical conduit of introspection.

Yet, this is a unique epistemological status in the history of scientific thought. Do we believe that the human brain can be understood by merely observing it from the outside? Certainly not! We estimate that it must undergo the scalpel of neuroanatomy, electrode stimulation, lesion neuropsychology, or computer modelling. Why should this not also apply to the intimate understanding of our lived consciousness? True, consciousness is invisible, accessed by none but the concerned subject, but the exclusivity of this access does not mean it is immediate. I will suggest in chapter ... that lived consciousness, far from falling under immediate observation, requires experiments to be conducted; that it relies, as it were, on concepts and knowledge surpassing what is found directly within it. But for now, I would like to turn my attention to other axiomatic systems concerning phenomenological experience and see whether they too contribute to this somewhat naive conception of consciousness.

15.3 Comparison with Other Axiomatic Systems (Aleksander & Dunmall)

Tononi is not the first to attempt capturing the burning sun of consciousness through axiomatic means, and he certainly will not be the last to singe his fingertips in the endeavour. His contemporaries, Igor Aleksander and Barry Dunmall, in a 2003 article titled *Axioms and Tests for the Presence of Minimal Consciousness in Agents*, raise the question of whether there exists a set of systematic principles to determine if a subject (whether biological or artificial) is indeed endowed with phenomenological consciousness. Quite interestingly, they propose a set of five axioms rather akin to those of Tononi. I wish to examine these axioms and the underlying approach, then compare them with what has been said concerning the Integrated Information Theory 4.0.

Right from the start of their article, Igor Aleksander and his co-author situate their approach within a realistic framework, positing that consciousness is the capacity of an agent to apprehend a sensible world S that lies before it and to evolve within it. This sensible world, in turn, comprises events or changes susceptible of being observed. Each event can be broken down into a set of minimal elementary events: one might think of the state of physical particles, the colour of space points, or the position of objects perceived by the senses. The authors do not specify precisely what they have in mind, but they use the example of a television screen composed of pixels. They thus base their approach on a reductionist, representational ontology, where the world is made up of a sum of irreducible units representable through the senses. In this context, understanding the whole equates to thoroughly understanding the relationship among the parts. Furthermore, this excludes the presence of non-representational states of consciousness, such as abstract ideas (justice, truth...), empty concepts (paradoxes), or ideas whose meanings exceed the representation one may form of them (the idea of infinity).

Hence, the sensible world S is potentially an infinite, yet enumerable, set of elementary parts $S_1, S_2, ...$ When a change occurs, even infinitesimally, within one of these parts, the subject, to perceive it, must himself exhibit a change in his internal state: neuronal activity alters, the variables in an artificial machine change values, etc. In other words, a mapping exists between the mind and the sensible world. This correspondence is not necessarily linear; multiple components of the mind may very well change in reaction to a single alteration in the sensible world, just as it is possible for multiple world events to be captured by a single change of an internal variable. However, the relationship must be injective: two identical neuronal states cannot coincide with the same world state. In contrast, two world states can be reflected by the same neuronal state, but in this case, it means that these

world states are not distinguished by the subject. To illustrate, we might think of a colour-blind patient incapable of perceiving a change of colour.

In addition to representing the sensible world, the subject, according to Igor Aleksander, possesses the capacity to recall states of the world or to generate similar sensations from nothing as this world has imprinted upon him. He can deploy this capacity for imagination to guide his attention, anticipate the unfolding of events, navigate the turbulent stream of events. These internal state changes are accompanied by affects, sometimes positive, sometimes negative, steering him sometimes this way, sometimes that, providing sort of a compass. These considerations are summarised by the following set of five axioms:

> **Axiom 1 (Representation):** The subject possesses perceptual states that represent parts of S.
>
> **Axiom 2 (Imagination):** The subject possesses imaginative states that remind him of parts of S, or create sensations similar to S.
>
> **Axiom 3 (Attention):** The subject is capable of selecting the parts of S he wishes to represent or imagine.
>
> **Axiom 4 (Planning):** The subject can at least partially control the sequence of imaginative states to plan actions.
>
> **Axiom 5 (Emotions):** The subject possesses affective states enabling him to evaluate the planned actions and determine the subsequent action.

In comparison to what has been laid out in terms of Tononi's theory, Aleksander and Dunmall's approach merits attention for initiating a mereology: the precise relationship between the whole and the parts is described, potentially explaining how consciousness can be both the entirety of perception (S) and the perception of singular parts within the whole ($S_1, S_2...$). However, this mereology is rather rudimentary. It assumes, on the one hand, that the sensible world takes part in the perceiving subject (realism), but also that consciousness is a product of neuronal activity (monism). Lastly, and this is perhaps the most questionable assumption, it is presupposed that the brain, like the environment, can be decomposed into indivisible units (reductionism).

These three assumptions just laid out align with Tononi's methodological principles, particularly realism (principle 2), operational physicalism (principle 3), and operational reductionism (principle 4). One might even be tempted to believe that Tononi conducts, as it were, a more systematic

exposition of Aleksander's intuitions. Nonetheless, I would like to criticise these principles and demonstrate the kind of issues they lead to.

To begin with, let's observe there is a distinction between structural analogy, which establishes an identification link between the mind's structure and that of neurons, and pure monism, which states that consciousness is a product of the brain. It is one matter to say that the brain, or a part of the brain, is a necessary condition for certain cognitive functions; however, this does not imply that consciousness is therefore "located" within neuronal tissues. Would it occur to us that sight is located in the eye? Certainly, the eye participates in the phenomenon of vision, certainly sight fades when the eyelids close, yet it is possible to have two functional eyes and see nothing. Vision involves many more organs than the ocular spheres, from the optic nerve to the cerebellum, through the lateral geniculate body, the ventrolateral path, or the dorsolateral path, so much so it is hard to locate it naively in a specific region of human anatomy.

Moreover, when we speak of "sight," we presume it to be a well-defined concept, though it encompasses itself a vast range of distinctions. The vision of fundamental colours, motion detection, shape detection, night vision, saccades, etc. Each distinction made within the concept of vision can also correspond to a functional distinction in various brain organs. Hence, the brain is broken down into ever more precise regions, whose interactions increase as we study it. But saying that vision reduces to a set of elementary components (colours, shapes...) corresponding, pairwise, to elementary components of human organs (retina cones, ventrolateral path...), is a mere hypothesis. Not just a hypothesis about human anatomy or about sight, but a hypothesis about the nature of language.

Why should Nature have wanted to clearly distinguish within our brain organs the function of sight from the sense of touch? Or sight from emotion perception? Humanly experienced phenomena are complex and disparate experiences, partially similar and partially distinct from each other, to which the concepts we use attempt to provide cohesion. Yet it is challenging to isolate a pure act of sight from a pure act of sensation. These merge so intimately within us that, without concepts, we might not have distinguished the sight of a luscious fruit from the hunger it incites—and thus would not have searched for distinct brain zones for the sight of a general object and for the processing of general emotions.

The concepts of our language draw qualitative boundaries between experiences that might ultimately rely on neighbouring anatomical mechanisms. The map of brain phenomena reading, if it were written by nature, would it be intelligible in human language? And wouldn't a being endowed with a superior intelligence have been able to do without concepts themselves?

Later in this thesis, I will demonstrate that the concepts we handle, including "sight" and "brain," require reduction to elementary units (neurons, space points, colours...) to be made intelligible, understood together

in their relations; so much so that the entirety of human knowledge is a vast mirror game made of comparisons, approximations, distinctions, and analogies, giving us the conviction that a direct link exists between the words used and the phenomena at play in nature. However, I will argue for a holistic vision, within which these units are ultimately merely abstractions, admittedly necessary but reductive, of phenomena considered in their interdependence. Thus, an alternative exists to the operational reductionism defended by Tononi, Aleksander, and Dunmall.

This complex relationship existing between the structure of language, the structure of the brain, and the structure of the observable world may contribute to the fact that neurosciences, taking empirical results seriously, have abandoned the purely localisationist conception, which consisted in pairing each cognitive function (vision, language, memory...) with a part of the brain (cerebellum, Broca's area, hypothalamus...)? Since the studies conducted by Pierre Jean-Marie Flourens on how pigeons, after a brain lesion, restructured the functioning of their brain areas, or Hughlings Jackson's work on epilepsy, neuronal plasticity is acknowledged. Nerve cells have the power to create associations among cognitive functions, motor functions, and certain types of stimuli.

Neuronal plasticity invalidates the attempt to map one-to-one each component of the "vision" phenomenon to a human organ, but the question can be posed whether the mind is located throughout the entire brain? In specific brain structures? Throughout the entire body, understood as a psychophysical unit? In the interaction of the individual with the environment? In nature entirely, akin to a German Romantic panvitalism? Or perhaps the very notion of location, applied to the concept of consciousness, gradually loses its meaning? This question is none other than that of the materiality of the mind. If it is something material, where is it? And why does my mind have no access to the phenomena presumably occurring in the minds of others? Or to phenomena occurring elsewhere in the environment? And if the mind is not material, what is its nature?

15.4 Physicalist Postulates

Tononi asserts that consciousness, far from residing within a static set of neurons, is a process of information processing. He introduces the hypothesis of a dynamic core within cerebral processes that coincides with the conscious state. This core generally includes the thalamocortical system but can expand, contract, move within the brain, or even split or vanish. Peripheral neurons might hence participate in this highly integrated information process. "We predict," writes Tononi, "that the neuronal complexity [of the dynamic core] should be much higher during wakefulness and REM sleep than during the deep stages of slow-wave sleep, and that it should be

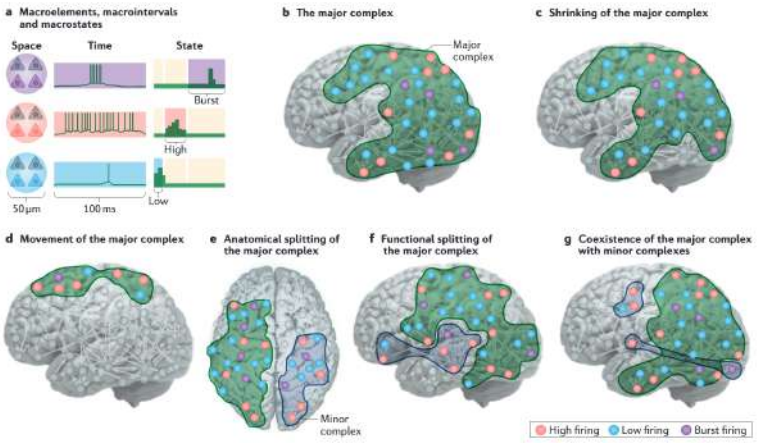


Figure 15.2: Source: TONONI G. et al., *Integrated information theory: From consciousness to its physical substrate*, in Nature Reviews Neuroscience, May 2016, p. 7.

extremely low during epilepsy despite the overall increase in brain activity[21]."

The question then becomes to determine the criteria for isolating this dynamic core, also known as the *physical substrate of consciousness* (PSC). Ideally, this should be an approach open to empirical corroboration or refutation. The general strategy involves associating each of the five axioms of phenomenological experience with a mathematical criterion, called a *physicalist postulate*. Neurons are modelled as a set of nodes within a graph, each capable of being in one of two states at any moment: "active" or "inactive". Two nodes are connected if there is a causal effect between the neurons they represent; in other words, if the state of one neuron statistically influences the state of another.

Mathematically, the probability that a certain neuron u is activated depends on the weighted sum of activations of the incoming neurons $u_1, u_2, ..., u_n$ according to a logistic function:

$$\frac{1}{1 + e^{-k \sum_{i=1}^{n} w_i u_i}}$$

where

- Each incoming neuron u_i is either in the active state ($u_i = 1$) or in the inactive state ($u_i = 0$).

[21] *Consciousness and Complexity*, p. 1850

- The weight $0 \leq w_i \leq 1$ represents the strength of the connexion linking neuron u_i to neuron u.
- The sum of the weights $\sum_{i=0}^{n} w_i = 1$.
- The parameter k represents the slope of the logistic function. In the examples considered by Tononi, it is often equal to 4.

This function simply captures the fact that the activation of an isolated neuron is given by a sum of Poisson processes, that is, binary variables that can be in a certain state with a given probability (active neuron) or in another state (inactive neuron). Therefore, a neuron's activation probability approaches 1 if all its neighbours are active, and it equals 1/2 when none of them are.

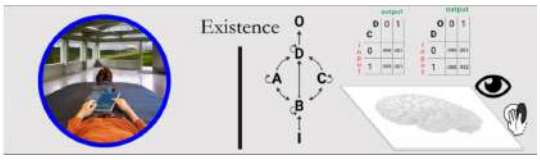
From this equation, it is possible to derive the state transition matrix $\mathcal{T}_U$, which describes at every moment the probability that the system, from a given state s, evolves to any state $u*$. By "state" of the system, we mean the assignment for each neuron of a value equal to 1 (if the neuron is active) or 0 (if not). The matrix appears as a table assigning each given state the probability of evolving to another state.

The probability matrix allows constructing several quantitative tools aiming to measure either the amount of raw information (entropy), the acquired information held by a part of the system when controlling the value of other parts (conditional entropy), the statistical distance between two neurons (relative entropy), or the statistical independence of a group of neurons relative to others (mutual information). By skilfully manipulating these tools, Tononi strives to construct, for each phenomenological axiom of consciousness, an empirical criterion. I will limit myself here to presenting the theoretical ideas[22].

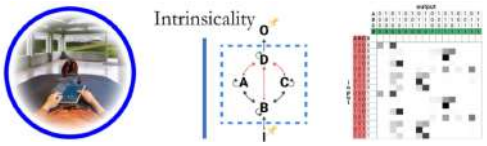
- **Existence:** My experience exists, immediately and undeniably. From this regularity, I infer the existence of a world independent of me. I call this "physical" if I can operationally observe (the eye) and manipulate (the hand) a substrate (e.g., a brain) whose units can "make a difference and be made different". This yields conditional probabilities that capture its causal power, summarised by a substrate model.

- **Intrinsicability:** Just as my experience exists for itself, the physical substrate of consciousness (PSC) must have causal power for itself.

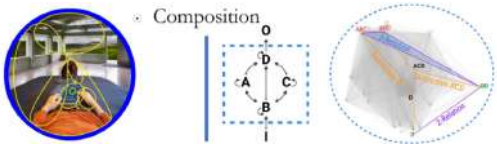
[22] These postulates are formulated in all of Tononi's articles, with variations in precision, technicality, or philosophical argumentation. I am using the version from the article *Only what exists can cause: An intrinsic view of free will*, which seems to me a good compromise between clarity, brevity, and formalism.



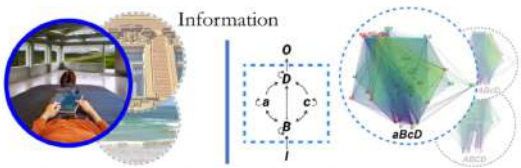
For instance, in ABCD, D is affected by A, B, and C, and affects itself. The dashed blue line indicates a candidate PSC whose causal powers can be analysed as a transition probability matrix (TPM), with fixed background conditions (I, O).



- **Composition:** Just as my experience is structured, the cause-effect power of the PSC is composed of distinctions and causal relations that form a cause-effect structure (blurred). The distinctions (orange) include a cause (red) and an effect (green) linked by a mechanism (black). The distinctions are connected by relations (purple, blue) if their causes and/or effects overlap.

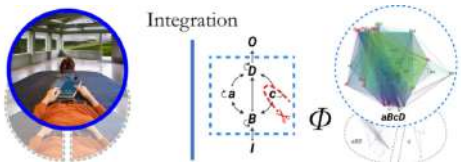


- **Information:** Just as my experience is specific, the cause-effect power deployed from a PSC in a specific state (aBcD – uppercase ON, lowercase OFF) is a cause-effect structure with specific cause-effect states. And just as my experience, being specific, can be differentiated from countless other experiences, the cause-effect structure specified by other PSC states (e.g., ABcD and ABCD) can be as well.

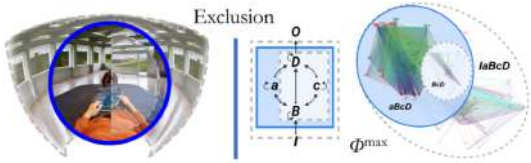


- **Integration:** Just as my experience is unitary, the causal power of the PSC must be unitary. If this is the case, one cannot partition this

substrate without losing intrinsic information. Simply put, the sum of the information contained in the parts is not equal to the information of the whole.



- **Exclusion:** Just as my experience is definite, the cause-effect power of the PSC must be definite. According to the principle of maximal existence (see text), the boundary of the PSC is the set of units, on a given substrate, that exists the most (here, aBcD, which has a maximal Φ), to the exclusion of all other overlapping sets (e.g., BcD and IaBcD).



15.5 Comparison with Shannon Information

Though the notion of information plays a central role in Integrated Information Theory (IIT), it differs distinctively from the one developed in Shannon's communication theory. The following passage aims to briefly distinguish between the two paradigms from an epistemological standpoint. It is assumed that the reader has at least a general knowledge of Shannon's theory. Otherwise, I would advise skipping this section or referring to Appendix **??**.

In Shannon's information theory, information quantifies the precision with which input signals can be decoded by knowing the output signals transmitted through a channel (noise). Mutual information represents the average statistical dependency between two sets of variables, such as inputs and outputs of a channel. In general, this depends on the entropy of the inputs and the ability of the inputs to predict the outputs. The (self)-information contained in a single state of a system is simply the negative logarithm of its probability of occurrence. As Shannon clearly acknowledged, information in information theory is dissociated from meaning: the sender and receiver are assumed to understand the meaning, and the focus is solely on the reliability and efficiency of transmission through a channel. Though

Shannon information can be evaluated between past and present states of the same system, it is always assessed from the extrinsic viewpoint of an observer assessing the statistical dependency between inputs and outputs. Thus, one could term Shannon information as extrinsic information.

In contrast, in IIT, information is intrinsic; it is evaluated from the intrinsic perspective of a system in terms of differences significant to it. Intrinsic information is causal and must be evaluated by perturbing a set of elements in all possible ways, not simply by observing them. It is also compositional, as different combinations of elements within a system can simultaneously specify different constraints. Crucially, in IIT, information must also be integrated: if the partitioning of a system makes no difference to it, there is no proper system. Furthermore, IIT considers only information that is maximal and exclusive, as, from the intrinsic viewpoint, there can be only one set of causes. None of these requirements matter from the extrinsic perspective of an observer measuring the global statistical dependency between inputs and outputs.

It should be noted that within the Shannon information framework, the difference between two probability distributions is evaluated using the Kullback-Leibler divergence (KLD), which measures the loss of Shannon information from the viewpoint of an external observer if one probability distribution is approximated by another. In the case where a distribution is evaluated in relation to maximum entropy (all states equally likely), the KLD measures the reduction of uncertainty. As information in IIT aims to capture differences that are significant from the intrinsic viewpoint of a system, it uses a true metric (EMD) sensitive to the relative distance between system states.

Most importantly, classical information theory does not deal with meaning, only with how messages are communicated and stored. In IIT, however, information is meaning: more specifically, integrated intrinsic information forms a maximally irreducible cause-effect structure (MICS). This integrated information structure captures all constraints on a complex's past and future as determined by its mechanisms' state. Thus, it is a form in conceptual space, not a message sent through a channel. Consequently, meaning resides within the MICS itself, and the meaning of each individual concept within the MICS is self-generated, self-referential, and holistic: it is constructed by the complex's elements, on the complex's elements, and in the context provided by other concepts within the same MICS.

For example, consider the firing of a facial neuron deeply situated in the brain and part of the main complex. From an extrinsic observer's perspective, such a facial neuron can certainly convey extrinsic Shannon information about certain environmental events, given its firing is strongly correlated with the presence of faces. From an intrinsic perspective, however, the firing of this neuron is meaningful due to how it modifies the shape of the MICS generated by the main complex (which it would modify differently

if silent). Consequently, dreams hold as much meaning as waking experiences: the firing of a facial neuron in a dream specifies the same intrinsic information even if it conveys no extrinsic environmental information.

Ultimately, of course, circuits generating meaning emerge, develop, and refine through a long process of evolution, neural development, and learning under the selective pressure of a complex environment. According to IIT, the goal of these circuits is not so much to process extrinsic environmental information. Instead, it is to generate integrated information structures that "fit" environmental regularities so that the "right dream" occurs at the right time[23]. This fit, in turn, may drive the evolution of consciousness.

15.6 Ontology: Between Eleaticism and Spinozism

A primary objection to Tononi's theory questions the range of conscious beings. Once consciousness is defined as a set of properties concerning information, there is no *a priori* reason to exclude non-living entities like the planet as a whole, stars, oceanographic processes, or even a mere heap of sand.

Tononi seems to anticipate this objection. In his essay on free will, published in 2022 on arXiv[24], he specifies that most neuroscientists frame their reflections within an extrinsic substantive ontology, while he defends an intrinsic power ontology. That is, according to the dominant thought, what exists are aggregates of matter, especially atomic units. Thus, mental reality is produced by neuron articulation. Tononi, however, proposes an intrinsic power ontology, where consciousness, understood as a maximal information system ϕ, is primary. All other things, including neurons, exist in relation to it extrinsically. They exist only for it, and not in themselves.

Furthermore, he distinguishes between Φ systems that are relatively maximal and those that are absolutely maximal. The human organism, for example, presents a relative maximum. The system comprising all our organs is more integrated and differentiated than a system with only some of them or a system including external objects. Our body thus presents a relative maximum. However, it is not comparable to the degree of integration found in the dynamic core of neural activity, which alone corresponds to a fully-fledged conscious activity—i.e., an intrinsic being.

> *Bodies and organs, tables and rocks, stars and planets, and many other things that seem to hang together well are likely to un-*

[23] See notably Tononi G., *Integrated information theory of consciousness: an updated account*, Arch Ital Biol 150: 56-90, 2012, and also 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.

[24] Tononi G. et al., *Only what exists can cause: An intrinsic view on free will*, arXiv, June 2022

> *fold into extrinsic entities. Many objects of study in the special sciences, such as molecules, organelles, cells, organs, organ systems, organisms, groups, societies, and the like, which have proven to be good ways of carving nature at its joints, are also likely to correspond to relative maxima of* Φ*. No matter how well extrinsic entities hang together, they do not exist for themselves. They only exist vicariously, from the perspective of some intrinsic entity, and so they do not truly exist*[25].

The distinction between relative and absolute maxima is not mathematically clear. It may refer to the analytical difference between a local maximum and a global maximum. The first is a point x whose image is greater than all other points in the neighbourhood, while the second is a point whose image is greater than all others in the domain.

However, I believe it would be more accurate to adopt a Spinozist reading of Tononi's words. What exists maximally is that which can have the most causal effects on other information systems and be affected by them in return. Integrated Information Theory also adopts the Eleatic principle, as formulated in Plato's *Sophist*[26]: "I say that anything which possesses any sort of power, either to produce a change or be affected even by the least cause, however slight, has real existence: for I claim that entities are nothing other than power." A similar principle is expressed by Buddhist philosopher Dharmakirti[27]: "Anything that has causal powers is truly existent." It should be noted that the Eleatic principle is expressed as a disjunction (either cause something or be affected by something), while IIT's ontological principle is a conjunction (cause and be affected).

In other words, the essence of being is the capacity to make a difference in the order of things. An existing system is one in which the information held impacts the information about the rest of the world and vice versa, without the stochastic behaviour of one being entirely predictable from the other. A maximum of information might correspond to something that cannot be conceived otherwise than by itself, not by virtue of something else. It is impossible to know an individual's consciousness without observing the dynamic neuronal core sustaining the phenomenon. Conversely, the things of the world, which are commonly considered material, are determined by consciousness much more than they determine it. Such an interpretation would align Tononi's ontology with Spinozism, especially with the axioms 1

[25] Tononi G. et al., *Only what exists can cause: An intrinsic view on free will*, arXiv, June 2022, p. 10.

[26] Cooper J. Plato: Complete Works. Hackett; 1997, cited in Tononi's article *Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms.*

[27] Tillemans T, *Dharmakirti*, in Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Spring 2021 ed. Metaphysics Research Lab, Stanford University; 2021, cited from the same Tononi article.

and 2 of Ethics, which respectively assert "Everything that is within itself [intrinsic existence] or in something else [extrinsic existence]" and "What cannot be conceived through something else must be conceived through itself."

However, the beings discussed in Tononi's theory are not substances as in Spinoza, but processes. Moreover, the information process, unfolding in physical space, and the mental phenomenon of consciousness are held to be one and the same; whereas Spinoza sees them as different attributes of the one substance. Thus, it would not be correct to say that Integrated Information Theory is a form of Spinozism, though similarities can be noted.

Among these similarities is Tononi's idea that there should be no dualism, no multiple categories of being, such as mind and matter, or thought and space, but rather that both are simply an organisation of information. He defends a sort of neutral monism where consciousness is the reference centre: "because I actually exist—as a large intrinsic entity—the neurons of my PSC cannot also exist. They cannot exist as constituents of my PSC, because what actually exists is not a substrate as such but the Φ-structure expressing its causal powers. And they cannot exist as small intrinsic entities in their own right because if they specify a large intrinsic entity, they cannot also specify a set of smaller entities. Moreover, because my alternatives, reasons, and decisions exist within my experience—as sub-structures within an intrinsic entity—the neuronal substrates of alternatives, reasons, and decisions cannot also exist. If this picture is correct, it leaves no room for emergence or dualism of any sort.[28]."

It should be noted, before concluding, that when the brain (or any complex natural or artificial system) is in a particular state, the quantity of integrated information Φ depends not only on the states activated in the system but also on states that are not activated but could be. Silent neurons thus contribute as much to the conscious state as active neurons do. In other words, consciousness is a product of integrated activity in the network and is measured by the integration of information, a property defined in graph structures. Knowing that a system is not in a certain state provides the same amount of information, all else equal, as knowing it is in that state.[29].

[28] Tononi G. et al., *Only what exists can cause: An intrinsic view on free will*, arXiv, June 2022, p. 10.

[29] More precisely, entropy, measuring the quantity of information, depends logarithmically on the probability of the state being activated. If the state corresponding to the concept "black", for instance, has a 20% chance of activating, knowing it is activated provides more information than knowing it is not.

15.7 Metaphysical Chiasms and Epistemic Principles

To conclude this chapter, Integrated Information Theory has the merit of showing how a scientific theory, expressed mathematically and objectively from a third-person perspective, can rely on an introspective theory of consciousness that seriously considers phenomenological experience, even making it the criterion of reality. Interiority becomes, in some sense, more real than exteriority, though the former can be expressed in the mathematical terms of information theory. Thus, Tononi and his colleagues' theory is closest to a dual-aspect monism or a neuroscientific Spinozism. However, there are some epistemological reservations to express.

First, the relationship between mind and matter is considered an identity relationship. Tononi does not accept the presence of an inexplicable dimension within phenomenological experience. On the contrary, all aspects of consciousness methodically lead to physical postulates. Conversely, the physical postulates are but the objective expression of these conscious properties. Ultimately, interiority and exteriority are two equivalent levels of explanation. Tononi even speaks of an operational reductionism: to theorise the world and identify information processes, it is necessary to divide this world into elementary units, but these units have no ontological reality *a priori*. The neurons discussed in relation to the brain and its mechanisms, for example, are essential units to scientific models, but they do not exist in themselves, apart from a prior consciousness. The mind is not reduced to the activity of a given set of neurons but stirs sometimes one part of the brain, sometimes another.

This strict identity between mind and matter conceals an externalist metaphysical chiasm. It is assumed that mental reality, in all its dimensions, must have a tangible expression in the extended space before us. However, Tononi's definition of consciousness relies on a distinction between an absolute maximum and a relative maximum of information that does not clearly identify why neuronal processes can be termed conscious and not the Earth system, ocean currents, or the union of several brains. It is indeed possible, theoretically, to invalidate or validate Tononi's theory experimentally, but this would, in practice, require highly complex measurements. It would need, no more no less, the manipulation of all neurons' activity to see which neural subset is most statistically independent. Then, one would have to compare the information presented by other equally complex physical systems. The conviction that one arrives, by this means, at the conclusion that only the dynamic neuronal core presents an absolute maximum of information is akin to an act of faith.

One might argue, however, that science, to advance, needs to be based on conjectures and that Tononi's, as difficult as it might be to verify, is

no more astonishing than many postulates happily reigning in the contemporary physics paradigm. I willingly concede this point and I am ready to admit that Tononi's theory could receive positive confirmation. In this case, however, it would mean that one possesses a complete representation of consciousness. Yet any representation, for first-hand experienced consciousness, is but a simple act of consciousness among others: it's an idea, a sensation, a sophisticated mental content indeed, but which exists *within* a prior consciousness. Therefore, a subject trying to understand consciousness with this mathematical representation cannot shake off the firmly anchored feeling that the mathematical theory does not capture the fact of being conscious in the first person.

This paradox, as I stated in the exposition of my thesis's general principles (hypothesis 2), is related to the idea of an infinite in actuality. Consciousness, regardless of the idea one has of it, always seems to overflow the representation. According to me, one must admit there is something unrepresentable within consciousness. Since all space-bound things are representable under geometric forms, this means consciousness cannot be strictly localised in space. I do not mean that science cannot speak about consciousness. It is capable of determining the boundaries of consciousness, i.e., a set of necessary but not sufficient conditions for the observable expression of consciousness: such as having a brain, being alive, etc. These conditions, however, are not absolute truths, but plausible correlations, admitting that consciousness expresses itself through space, without showing that it is, for that matter, entirely apprehensible in spatial terms; like a cylinder suspended above a sheet of paper casts its shadow on that sheet without being contained within the plane of the sheet.

A second critique of Integrated Information Theory concerns its somewhat static and naive view of phenomenological experience. Tononi states in several places that consciousness is a process, not a thing—and that is why he does not seek to localise it in brain parts, but identifies it with a dynamic information system. Yet when he describes phenomenological consciousness, he seems to forget this processual dimension of consciousness somewhat. None of his five axioms of lived experience makes reference to dynamic law.

Unlike Newtonian kinematics, which describes equations of motion, consciousness in Tononi appears to be given entirely in the moment. There is the thought that phenomenological experience is something transparent, immediate, and total. When a subject turns their gaze inward, they contemplate their mental life as if it were a landscape. This can be decomposed into hills, trees, impetuous rivers swept by a silent wind, but this naturalistic anatomy does not disclose the secret of movement stirring impressions, thoughts, sensations, and connecting them to one another. Indeed, consciousness is not within the landscape; it is the eye suspended over all the *possible* landscapes that form the unbroken thread of life. It is the wide-open pupil of Victor

Hugo's poem, which continually watches Cain, reminding him of his existence as a subject trapped in the world:

> *Lifting his head, deep in the dark skies,*
> *He saw a great eye wide open in the shadow,*
> *And gazing fixedly at him.*

To enrich Tononi's theory and make it compatible with a dual-aspect monism, it is necessary to embark on formulating an introspective theory of consciousness, starting from experience in all its complexity and, without simplifying it, identifying several universal principles. Such an approach must account for the apparent paradox between axiom 3 (integration) and axiom 5 (composition) of Integrated Information Theory, which state that consciousness is a unit, a whole given as a block in mental life, yet this whole is decomposable into parts. The consciousness theory I aim to construct must, therefore, be a mereological science, studying the articulation between different mental contents and consciousness as a whole.

How does it happen that when consciousness focuses in turn on the river, the wind, the bending willows, it doesn't split into consciousness of the river, consciousness of the wind, consciousness of the willows, but remains the same consciousness set upon a unified landscape? Why, conversely, must the landscape be divided into so many objects, and not be comprehended as an irreducible ensemble? No two trees are alike, and no two rivers are identical. Why does human consciousness feel compelled to divide the world into concepts? These remain unanswered questions.

Key Takeaways from this Chapter

In the tradition of Gerald Edelman, with whom he has published several articles, Giulio Tononi advocates that consciousness is not located within a group of neurons but within the brain's reentrant processes.

He seeks to develop mathematical criteria for identifying the dynamic neuronal core involved in conscious activity. This core is not fixed: it is neither, *stricto sensu*, the thalamocortical cortex, nor the limbic system, prefrontal lobe, or default mode network but neuronal activity that can expand, contract, move, disappear momentarily, or split.

Methodologically, Tononi begins with the phenomenological experience of consciousness, described using five axioms, matched to six physicalist postulates. He then constructs a measurement of information called integrated information.

I suggested that Tononi's approach, allowing for a Spinozist reading, approximates a dual-aspect monism, but it attracts two objections:

1. The mental is thought to be identical to the material: nothing in conscious life cannot, according to Tononi, be represented as information processes. This implies that the body-mind relationship is one of strict identity, not paradoxical complementarity.
2. Mental life is portrayed as a static landscape, not dynamic. In other words, consciousness falls under a regime of immediate, transparent, and total introspection. No mechanisms are provided explaining why the entirety of conscious perception endlessly divides into differentiated objects.

Filling these two gaps will be the aim of the third section of my thesis, titled "Mereological Theory of Consciousness."

Chapter 16

Foundations of a Mereological Theory of Mind

If I might summarise the findings achieved so far, I believe I have suggested that dual-aspect monism holds potential within contemporary neuroscientific theories. On one hand, there's a school of thought, championed by Henri Atlan[1] and Antonio Damasio[2], rooted in a Spinozist interpretation of the body-mind relationship. On the other hand, I have demonstrated that authors like Gerald Edelman and Guilio Tononi propose a theory of consciousness establishing structural analogies between mental and physical spaces. Although they did not explicitly express it, their approach can be deepened to support a bi-perspectivist view of body and mind inspired by the Jung-Pauli hypothesis.

Despite these promising outcomes, I have identified certain blind spots within the neuroscientific framework. Notably, with Tononi, there is an underpinning computationalist ontology, which equates consciousness, and cognition more broadly, to information processing. I have highlighted that such an approach faces fundamental challenges: chiefly, it struggles to draw a clear line between what should, and should not, be conscious. Despite a mathematical criterion offered by the Integrated Information Theory, ensuring that this criterion doesn't become trivial (all complex processes are conscious) or overly restrictive (nothing is conscious) in practice is difficult. Currently, it remains hard to verify, requiring intricate calculations of integrated information amounts within the brain's subcomponents.

Furthermore, Tononi's theory presents a somewhat limited vision of phenomenological experience, based on the presumptions that (1) introspection is immediate, (2) consciousness is at once unified yet composed of parts, (3) consciousness can be captured through static properties without detailing internal dynamics, and (4) consciousness merely processes environmental

[1] Atlan Henri, *Les étincelles de hasard*
[2] Damasio Antonio, *Spinoza avait raison*

information without actively shaping reality. Hence, in this section of my thesis, I aim to lay the groundwork for a methodological approach, termed *mereological*, which revisits these dimensions.

16.1 Critique of Cognitive Empiricism

While the exact number and nature of faculties within the realm of consciousness remain debatable, an implicit consensus agrees that impressions[3], ideas[4], sensorimotor schemata[5], perceptual similarity spaces[6], and other derivatives of the human mind arise from interaction with the sensible world. Our initially rudimentary cognitive tools gradually develop through interaction with reality, achieving the sophistication recognised in adulthood.

Empiricism pervades cognitive sciences so thoroughly that Susan Carrey, in her book *Core Cognition*, considers herself subversive when she writes: "I question a widely shared presupposition about the ontogenetic origins of human conceptuality: the notion that initial resources of human mental representations are confined to perceptual and sensorimotor primitives[7]." Her thesis claims humans are born with innate cognitive structures, which she calls, in line with Elizabeth Spelke, *core cognition*. Though not acquired through individual experience, this faculty is selected by evolution, maintaining the empiricist view that the mind shapes itself in the contact with reality. It simply shifts from individual, ontogenetic experience to species-wide, phylogenetic experience.

In the dominant cognitivist perspective, humans, with or without *core cognition*, are akin to blind creatures, cast by chance into our hostile world, tentatively seeking their path. Representations gradually accessed, from the earliest sensorimotor stages to the age of reason, allow navigation in this environment and acquisition of resources necessary for survival. Debate persists on how these representations are constituted—via biological selection? learning? ontogenetic schemata?—but regardless of the adopted view, concepts are shaped by the "struggle" of the individual, or their species, for survival. Emotions, within this paradigm, are often seen as sources of motivation[8], or as feedback enabling individuals to select objectives and assess accomplishments[9]. In many respects, humans are conceived as vast adaptive

[3] Hume

[4] Locke and 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. Translation is mine.

[8] Alain Damasio

[9] Igor Aleksander and Dunmall.

machines, acquiring useful skills in a continuous feedback process of trial, error, and correction.

Unsurprisingly, starting from such an epistemology leads to the conclusion that mind emerges from matter. Since adaptation to the tangible world has been perceived as a prerequisite for any structured mental activity, only a small step remains to deny the mind's autonomous existence. Our contemporaries are far from Berkeley's idealism: "As it is impossible for me to see or feel anything without having an actual sensation of it, it is equally impossible to conceive in my thoughts a sensible object distinct from the perception of it." The rampant eliminativism in today's philosophy, rejecting notions of mental states, intentions, will, and even consciousness, is more an origin of empiricism than its result. It is the coherent principle behind the idea that "knowledge comes from worldly experience." Any attempt to bring lived (or phenomenal) consciousness back into the scientific horizon, without challenging the pragmatic thought that mental life objects are shaped by the material world, risks significant difficulties.

Thus, to honour the dual-aspect monism hypothesis, it is crucial to offer a theory of the human mind that maintains "ontological neutrality." What do I mean? If we are to avoid the mind being reduced to matter—or matter to subjective perception—the theory I aim to construct must support two concurrent interpretations. It should be compatible with both the empirical reading, wherein the birth of concepts occurs in neuroanatomical development, and the idealist reading, where the world forms from an internal structuring of the conscious universe.

Notably, I used the term "idealist" but could have more aptly chosen "transcendental[10]." My theoretical approach does not claim the world exists solely within subjective perception; it states that the world's existence or non-existence is irrelevant to its purpose. As Laplace exclaimed to Napoleon, "I had no need, sir, for the hypothesis of God!"—so, my philosophical system, well constructed, should operate without explicit reference to the noumenal world. What then underpins it? The distinction between the perceiving subject and the perceptual object.

16.1.1 The Problem of the Genesis of Consciousness

From the standpoint of vibrant consciousness, there exists no intrinsic difference between perception and illusion, sensation and imagination, or truth and falsehood. All these conscious manifestations indeed reside within us, expressing our existence as subjects. Thus, for a patient with Charles Bon-

[10]The term transcendental refers to Kantian theory, economising on the noumenon (the thing-in-itself), or more precisely denying knowledge of things-in-themselves, without repudiating their existence

net syndrome[11], hallucinations resembling abstract paintings appear as real as the vision of their own body—even though aware of their hallucinatory nature.

Conscious experience doesn't *a priori* distinguish between perception and hallucination, though distinctions between real, false, tangible, illusory, and material versus imaginary are deeply embedded in daily experience. The real question is not: why are human mind concepts suited to the external world? Rather: how does the mind, in solipsistic isolation, become convinced that objects of thought exist independently of thoughts about them?

Explaining why consciousness intrinsically separates thought acts (conceiving something) from thought objects (the things conceived) illuminates the genesis of intentionality. This genesis, concerned solely with mental life mechanisms, remains ontologically neutral. Consider we possess an *intentional mechanism* that accounts for the consciousness-based subject-object distinction, demonstrating (1) it aligns with realism, and (2) it aligns with idealism.

For point (1), we attribute consciousness's mechanism development to real-world adaptation. It is certainly more advantageous for survival—without contradiction—to maintain the concept of an external world rather than being ensconced like a Leibnizian monad in consciousness's closed universe. For point (2), we argue it is the intentional consciousness mechanism itself fostering belief, though superfluous, in an external world's existence.

Advocating interpretation (2) doesn't entail impracticality in predicting future phenomena, engaging with stable and reasonable frameworks, or showing respect towards others[12]. Mental objects, mere illusions perhaps, are still governed by fundamental constancies. What practical difference exists between conceiving natural laws inscribed mathematically in nature or in consciousness? What tangible difference lies between believing one inhabits a material atomistic world or a dream?—a dream laden with objects and structures indistinct from consensual reality.

True, these ontological constructs impact significantly on questions like: What is life's meaning? What follows death? Or what defines good and evil? Fortunately, these inquiries are speculative, whereas consciousness mechanisms are factual matters—albeit not scientific, more accurately deemed existential or introspective facts[13]. Indeed, speculative questions are inher-

[11] Charles Bonnet syndrome refers to complex visual hallucinations occurring in elderly patients with visual deficits from aging (cataract, glaucoma, macular degeneration...) but no mental disorders.

[12] These objections are common in neuroscientific articles against realism. For instance, Tononi states without substantial argument: "We should assume that something exists (and persists) independently of our own experience. This is better than solipsism, which explains and predicts nothing." (*Op citation*, p. 6.)

[13] The term "introspective fact" might imply the consciousness mechanism's direct access through immediate mental inspection. Yet, "fact" does not connote "immediacy." Scientific truths, for instance, are undeniably facts, but there is nothing immediate in

ently independent from factual answers. I do not deny facts may sway belief one way or another, but they do not constrain it. Consequently, if the aim were asserting the consciousness existence among other species, prompting moral respect, upholding real world existence might prove beneficial; yet, in elucidating the consciousness mechanism, this real-world premise is unnecessary—and can prove detrimental.

Assuming or disregarding reality's existence, the knowledge theory I intend to explore considers consciousness a kind of receptacle—a vessel infused with thoughts, sensations, doubts, fears, pleasures; containing the totality *possible* of lived experiences. This consciousness is, indeed, *the* world in entirety. Beyond this circumscribed realm, nothing can be conceived, experienced, felt, touched, thought. Consequently, at a given moment t, the subject experiences being what they feel, see, think. If such sensations, perceptions, thoughts were unchanging, the subject would lack reason to believe in any difference between self and world. They'd exist in Parmenidean immobility, where all that is, is consciousness, and all consciousness is, is all.

Yet, as an experimental fact, consciousness content changes over time. The consciousness I have at one moment evolves by the next. This progression reveals the world unfixed as first perceived, always a step beyond acquired understanding. I once thought I saw a person roadside, only to find, like Diderot, it was a mere stump. I thought I faced a beggar in rags, only to realize, like Telemachus, it was my father Ulysses. Experiencing change, error, illusion distinguishes my conscious capacity from known, shifting things.

These reflections lead to hypotheses about consciousness: the distinction between perceiving subject and perceptual object is inconceivable outside temporal flow. In consciousness, time embodies mental content change. Husserl stated consciousness is always "consciousness of something." I argue consciousness is primarily consciousness "of something that changes." This is the *principle of differentiation*.

Moreover, I advocate for critically reviving consciousness's creative and imaginative prowess. World perception engenders consciousness's creative force. Far from a passive world receptacle, it actively participates, moment by moment, in forming mental content. Evidence of this active consciousness power lies in numerous perceptual illusions, particularly colour saturation illusions (see figure **??**).

I wish to envisage consciousness not within a purely materialist vista, demanding organism adaptation to environment, but within a dual-aspect monism supporting dual interpretations: consciousness, as seen via evolutionarily adapting organ to environmental constraints; also as a conscious-

devising hadron collisions to infer gravitational insights. The experimental setup contains an externality providing access. Similarly, this thesis contends that consciousness's introspective or existential facts require non-introspective concepts and theories.

ness experience, encompassing the entire perceived world within. Thus, it is both an object within the world, an insignificant part amidst surpassing totality, and the world in its entirety. The question arises: how might this consciousness be studied? Is it feasible to approach it as an object without succumbing to metaphysical chiasms highlighted by the neuroscientific paradigm? How to ensure introspection dialogues with universally scientific methodology?

16.1.2 Dual-Aspect Investigation Method

To draw parallels between mental function and scientific world conception, perhaps inspiration lies in grand physical principles. Cosmology delights in studying the Universe's evolution from the Big Bang, perceiving it contained in a pinhead, then outlining particle trajectories. Similarly, I wish to unearth consciousness's Big Bang; a fusion point from which matter spreads in physical space and mental phenomena in psychic space. Such a state, if it exists, should reconcile the contradictions between matter and thought, internal and external, subject and object. It should discern unity behind dual perspectives.

Ideally, this state should be investigated from a scientific view and a first-person introspective view, to align their perspectives. That is, our coveted consciousness Big Bang must transcend metaphysical chiasms. Initially, this seems unachievable—I sense reader scepticism looming like a Sword of Damocles. Let me reassure you promptly: isolating pure vibrant and pure dead consciousness is unfeasible. My earlier assertions remain. Yet, I propose it's possible to converge towards a point where chiasms fade. To articulate what this Big Bang might resemble, consider these queries: what remains of the brain when all stimulation tends to vanish? What remains of vibrant consciousness as it empties of content?

Viewed *scientifically*, a brain with negligible neuronal activity—or none—approaches *pure* dead consciousness. Nothing phenomenal, nothing lively remains. No thought, sensations, doubt. Only raw nerves. Bereft of mental states, the notorious scientific zombie paradox dissolves. I am uninterested in cryogenically frozen brains, rather in individuals consenting to experiences reducing cerebral stimulations. This isolative device could deprive them of light, scents, sounds, tactile sensations.

From a *philosophical* introspection perspective, a mind devoid of mental states—or idyllically inactive—would resemble a *pure* consciousness receptacle. Wasn't this the Ego's definition? A container without content, an aquarium devoid of water and fish. Achieving this, akin to consciousness absence, would require meditation quelling the mind. Thoughts, impressions, anxieties, soul turbulence must settle, allowing hypothetical mere presence's enigma.

If these two activities, introspective meditation and brain stimulus isola-

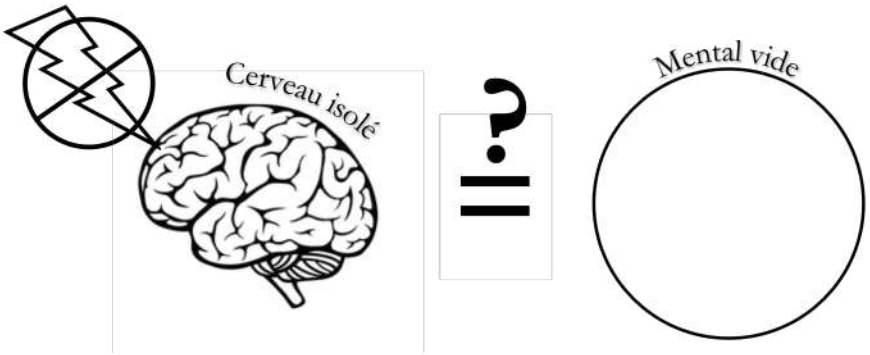


Figure 16.1: Is a brain isolated from all stimulation equivalent to an empty mind?

tion, occur together, vibrant consciousness and dead consciousness converge. At limit state, null activity brain and empty consciousness merge. The Ego becomes the brain entirety. Could one counter this resembles studying a deceased's brain? That it does little for consciousness phenomenon understanding? Admittedly, were cerebral demise reached. Yet, crucially, my approach interests itself in convergence process occurrence, avoiding complete experiment fulfilment. Could entire mental activity be suppressed? My goal is not committing mystical homicide!

If consciousness is, as argued, represented by a hypersphere, aligning Il and Je perspectives proves impossible. Rotate the four-dimensional object as you may, projected spheres in three-dimensional space never merge. Persistently, I refute Je equates with the entire brain, or consciousness merely brain phenomena. Near boundary state, expect a repulsive activity preventing Je-brain identification. This parallels confrontation with death, albeit harder studied.

From these ruminations, I conclude nearing void state means Either vivid consciousness images form, or brain activity emerges. Absent these, consciousness representation as a hypersphere warrants questioning, necessitating revision. However, if my prediction holds, it supports my theory. It's no proof—indeed, metaphysical proofs are elusive. Yet, it attests to my idea's viability. Venturing into speculation's perilous realm, this marks a noteworthy triumph. Next, my discourse explores a case study by researcher John Cunningham—describing it in the coming section—and highlights its essence akin to a genuine mystical experience. A laboratory mystique, one could muse.

16.2 The Isolated Brain Experiment

It is 1952, in a Washington laboratory. The place is immersed in a curious chiaroscuro. In one corner, the silhouette of a table strewn with creased papers emerges. There lies, haphazardly, treatises on neurology, psychoanalysis, and Hindu philosophy. A thin beam of light filters through a dusty window, cutting across the space before colliding with a metal chamber. Inside, water floats silently. A hatch allows entry, yet no sound escapes. This peculiar sarcophagus awaits a solitary man to embark on a strange journey. Equipped with a respirator mask specially designed to block light and a clinging suit, individuals take turns entering the tank. Their faces wear an odd mix of nervousness and scepticism. But upon emerging, their smiles are radiant. In this room at the National Institute of Mental Health, John Cunningham Lilly has just developed the first flotation tank, also known as an isolation tank.

Inside this unusual chamber, no taller than a man, the subject is deprived of all familiar external and cenesthetic reference points. Proprioception becomes challenging within the surrounding liquid. Touch and sight are significantly diminished. As for hearing, it reduces to a distant drumming. The individual, like a baby floating in amniotic fluid, drifts in space and time, devoid of any anchors or stimuli. In this context, it becomes impossible to cling to the cognitive and perceptive impressions that are foundational to the intimate feeling of existing in the world. But what becomes of the brain when stripped of all kinds of stimulation?

The majority of neurologists of the time, with psychoanalysts at the forefront, would not have hesitated to answer: the brain enters a state of rest akin to coma or biological death. In his essay *Beyond the Pleasure Principle*, Freud popularised the idea that in the absence of external forces, the psychic organism seeks to reduce both internal and external excitation as much as possible. According to this concept, named eloquently by his colleague Barbara Low as the Nirvana Principle, there was no reason to think that a brain isolated from environmental noise could retain biochemical activity. This intellectual concession is illustrated by Freud in his essay *Beyond the Pleasure Principle*:

> *"The conviction we have acquired that psychic life, perhaps nervous life in general, is dominated by the tendency towards reduction, constancy, and suppression of internal tension caused by excitations (by the Nirvana principle, to borrow the expression of Barbara Low), this conviction, we say, constitutes one of the most powerful reasons inclining us to believe in the existence of death instincts*[2]*."*

The Nirvana Principle naturally stems from Freud's desire to place psychoanalysis within a scientific methodology. The reason is clear: as men-

tioned in the opening chapter, science does not envision the mind as a *living consciousness*, but as a *dead consciousness*. In its view, the brain is a cranial box where signals are emitted in response to incoming stimuli. Thus, the scientific methodology consists of manipulating this black box of the psyche to study its reactions according to a rigorously established—and promising—protocol of cause and effect manipulation.

The notion that the brain could produce autonomous activity in the absence of all input was not easily conceived. Yet it is ironically amusing that Freud, as if wanting to provide fodder for my definition of *dead consciousness*, spoke precisely of *death* instincts to describe the psychic machine's tendency to return as quickly as possible to an inertial state. "The elementary living being," he writes further in his essay, "would very willingly have remained immovable from the beginning of its existence, asking for nothing more than to lead a uniform life in unchanging conditions."

According to Freud and his followers, the psyche is an inert organ yielding nothing to the Parmenidean Being[3]. It is a closed world, steeped in the desire to revert to its original state of unicellular being, which would have scarcely evolved without the interplay of foreign forces. It would soon die too, if not for the external environment and the cunning laws of evolution forcing it to take detours along the long path from birth to extinction. Far from an organism that, driven by its own force, grows despite mortal constraints, the brain of psychoanalysis is an organ seeking, in the labyrinth of life and sexual drives, a short cut to the peace of death.

In this somewhat morbid atmosphere, dominated by a stubborn esprit de corps, John Cunningham, with his isolation chamber, managed not only to bring an empirical dimension to the question of brain activity in the absence of stimuli but also to demonstrate to the initially sceptical scientific community that the Nirvana Principle needed revision. Subjects emerging from his lab describe a sense of inner peace, but the initial period of rest in the flotation tank is gradually replaced by hallucinatory activity worthy of the great ecstasies of Christian mysticism. Ultimately, it is as if the brain, abandoned to this state of radical homeostasis, took it upon itself to reproduce the stimuli from which the experimental setup had so skilfully cut it off.

Imaginary perceptions, dreamlike thoughts, and phantasmic desires then take the place of the usual perception of the world. The isolation tank functions much like Charles Bonnet Syndrome or the musical organ syndrome: when the senses of sight or hearing are diminished, some subjects, without brain lesions or mental disorders, may form complex perceptual or auditory hallucinations. Charles Bonnet, the first to describe this disorder, reported that his grandfather saw horse-drawn carriages, tapestries, figures, birds, and scaffold-like patterns reminiscent of the Hanging Gardens of Babylon[4].

What would a brain devoid of activity resemble? Perhaps the analogue of a landscape without sky or light, a gaping space, hauntingly void of any

vital breath to seed its pristine horizons. For the subject to exist as a *living consciousness*, and not merely as a being beckoned by death, it becomes imperative to populate their mental universe. Stasis in perception, a frozen image, would equate to the cessation of psychic existence—and perhaps herald a state akin to death? There are no notable differences, regarding the intensity of psychic activity, between the usual state of mind interacting with its environment and the hallucinatory state of a mind cut off from the world.

What a Copernican revolution this revival of neuronal cells becomes! In the traditional biological paradigm, the essential attribute of the brain, honed by years of evolution, was to identify, anticipate, and mitigate environmental threats. In this new approach, the notion that the mind's role is not to survive in the world, but to withstand the absence of the world, marks a breathtaking shift in perspective, shifting the focal point: the living being no longer solely uses consciousness to apprehend the world; but consciousness becomes a world in itself that the living seeks to sustain. We transition from a predominantly realistic ontology, where the world is real and the mind functions as its perception, to a more idealistic ontology, where the mind is real and the world merely its product.

Certainly, one could argue, and it is often done, that brain hallucinations have an evolutionary function. What does this mean? That the mind, when not subjected to environmental constraints, exhibits spontaneous activity enabling it to refine its models of reality. Sleep, dreams, and imagination, according to this biological worldview, serve as tools providing individuals with the means to simulate dangers and confront specific emotional contexts.

The brain might then resemble an idle machine. In our capitalist world, idle hands are generally frowned upon. If the brain has nothing to do, mustn't it find, on its own, ways to exercise its talents? And what is it better at, but preparing to face reality? The use of terms, more or less linked to computing vocabulary, often bolsters such instrumental explanations of dreams that indirectly equate the brain to a training machine within a virtual reality universe. "In many scientific circles, the belief that the brain is a computer remains widespread," said Gerald Edelman after winning the Nobel Prize in medicine[5].

However, it would be misguided to dismiss the brain-as-simulator theory due solely to terminological misunderstandings. Though risky to model the experiential object on technical tools we employ (like the computer), naivety in science is not a sin—and often preferable to scholastic sophistication! While one may doubt the brain as a computer, some credit must be given to the biological conception of the brain. We routinely experience unconscious learning in the sleep state. This occurs when crystallising knowledge repeated just before sleep or when faced with a seemingly unsolvable problem, we awaken with new hypotheses. Indeed, some top athletes use flotation tanks before important events. These help improve concentration,

presence, and confidence.

It is therefore entirely possible, even likely, that the evolutionary explanation is correct, but that doesn't make it the true one. Truth, in fact, not only requires sufficient reasoning; it also necessitates pursuing alternative answers. I believe that the essential question raised by John Cunningham's experiment is not *solely* whether the brain produces activity in the absence of stimulation—and if this results from evolutionary necessity. It also involves seizing the opportunity to explore, parallel to the brain question, the nature of the consciousness *I*.

To comprehend this aspect of the problem, which isn't strictly scientific but philosophical, one must consider the perspective of living consciousness. To it, the world is nothing but a phenomenon of consciousness whose existence outside our senses can be questioned. More precisely, it is indifferent whether this world exists beyond the perception of it, for the mind will never access anything beyond the perceptions themselves. A question like, "Does the brain produce activity outside of external stimuli?" bears little meaning. The real question is, "Is it possible to be a consciousness without mental content?"

In usual conscious experience, experienced outside an isolation tank, we are a mind filled with ideas, sensations, emotions, impressions—sometimes vivid, sometimes imperceptible—simultaneously blending sight, sound, taste, touch, and smell. Yet when we sever all ties with the world we knew and enter isolation, numerous contents are altered. Our perceptions, ideas, and sensations are no longer what they were. We might even fear that they vanish entirely.

Was it not plausible for John Cunningham's contemporaries to hypothetically consider that isolation might still the mind? Some inexperienced subjects do fall asleep in a flotation tank. Though sleep is usually populated with dreams, nothing forbids us from thinking, prior to experimentation, that the isolation tank might induce dreamless sleep—or a coma akin to an operating theatre. This would mean consciousness is suspended. It would be akin to a small death of the mind, perhaps analogous to the destiny our conscious path ultimately leads us to?

Note well, this question doesn't equate to asking whether there is brain activity. For the consciousness *I*, the objective brain does not exist; if it does, it is unobservable. We may at most possess the idea of the brain, but not the thing in itself. Thus, while the philosophical concern I outline certainly maintains an evident analogy with the scientific dimension of the question, its formulation cannot be substituted. This formulation consists of asking whether it's possible to be consciousness without anything to think, feel, or experience? And if, indeed, this must be excluded—as it is—by the isolation tank experience, it remains to explain why the mind exerts such effort, contrary to the Nirvana principle, to avert sensory extinction.

To answer this profound question, we must first ask what a living consciousness devoid of internal or external sensations—free of memories, dreams, and desires to explore—might resemble. This scenario is, so to speak, an

extrapolation of the isolation tank—an intellectual experiment from which we hope to gain insight into the nature of the mind. Moreover, it represents the ideal that practitioners of Hinduism, Jainism, and Buddhism aim towards when aspiring for Nirvana.

16.2.1 I think, therefore I link

The "I" is a mere consciousness accompanying all concepts.

Heidegger

Cunningham's experiment results demonstrate a repulsion between the poles of dead consciousness and living consciousness. It's as though emptying the mind of thoughts or the brain of stimuli equates to walking with a compass toward the Antarctic centre. The compass, specifically calibrated to distinguish North from South, spins erratically as we near the Earth's extremes. The needle seems caught in a storm, trembling, skipping, slipping. The dial suddenly shatters, yet we continue advancing towards the source of this curious magnetism.

What must be studied, as a daring sailor, is the calm following this storm. I wish to ponder what a consciousness truly emptied of its contents would resemble? An empty consciousness, as I've said, is unrepresentable, but this doesn't mean philosophical contemplation is impossible. On one hand, I can rely on the sensation experienced during limit consciousness experiences. On the other, I can attempt to extrapolate abstract ideas from there. My approach, I admit, resembles conducting a thought experiment where nothing can be thought. Empty consciousness is the explorer's dream continent. The horizon imagined by the cartographer. Unlike Columbus, the chance of reaching the Indies of the human mind is slim. Yet, I believe there's hope of framing the internal dynamics of consciousness conceptually. I have shown that a vitality principle directs mental activity. The psyche, according to laws yet to be elucidated, prefers hallucinating over returning to rest. Now, it's time to understand whether other principles can articulate alongside this first one?

I invite you to close your eyes for a moment. Imagine with me the flotation tank experience extends so completely that we've individually ruminated each of our thoughts, memories, and sensory impressions. That countless universes of dreams, phantasmagorias, gaping metaphysical and traumatic doubts are born, accompanied by their mystical cohorts of imaginative monstrosities; that countless immaculate spaces of thought, dream, and speculation die. What remains then of this spirit that finds nothing, neither within nor outside itself, to savour? Close your eyes and ponder.

Do you have your idea? For my part, I believe it will retain only its most

essential property: being conscious of something at any given moment. I have called this property the I, but let's be clear: the I doesn't denote being a particular person in the world, with a certain name, role, or history. The I is devoid of identity, memory, and narrative. It is merely the acknowledgment that, at this given moment, there is a consciousness phenomenon. Though the content of this consciousness constantly changes, the fact of being conscious remains constant over time.

I am the world's contour

The I is the invisible boundary of the mental world, supporting the latter without ever presenting itself to it. It's neither a drop of water nor an ocean current: it's the oceanic basin. A fish lost in this ocean, questioning its identity, won't even realise the element it floats within daily. Ideally, it should leave the water to acknowledge its existence and understand its nature. Yet outside the water, it dies immediately. That's why the I is so fundamental, but why, despite its omnipresence, no one can point to it. "This centre of the world, which I name the I" wrote Wittgenstein, "is the support of ethics[6]." This is, strictly speaking, the essence of what is called the Mystical. "It cannot be explained, it cannot be stated, but it shows itself[7]."

This I, whose portrait we sketch, appears in several philosophical traditions. It first appears in Descartes' exclamation: *cogito ergo sum*. The father of modern philosophy's approach is indeed analogous to what we've charted here. Though he didn't invent the isolation tank (he contented himself with the still contemplation of ceiling stains), he doubted his senses' reality with such acuity and eloquence that he convinced generations of thinkers that the subject's most essential property, that about which no doubt was possible, was the act of thinking—not merely thinking of a particular thing, but being a thinking substance[8].

The presence of this I is also found in the opening chapter of Hegel's *Phenomenology of Spirit*, titled *Sense Certainty*. The philosopher of Jena envisions the nascent spirit of humanity in an immediate relationship to knowledge. He names sense-certainty the capacity to conceive singular things, concrete *thises* and *thats*, like this house here, without being able to relate this contemplation to a concept. In this relationship, the I that conceives confounds itself to the act of sensation. It's this singularity that distinguishes, for Hegel, sense-certainty from perception. "This sense certainty," he writes, "presents itself as the richest and poorest truth. It speaks of what it knows only in this: *it is*; and its truth solely contains the *being* of the thing occupying it; consciousness, on its side, is in this certainty only as pure I, or again, I am merely pure *this one*, and the object is similarly only pure *this* [...] consciousness is I, nothing else, a pure this individual here [...][9]".

Another important source for our concern—but I'm not pretending to consider them all—lies in Hindu tradition. Particularly interesting is the reflection found in the oral teachings of Sri Ramana Maharshi, renowned for his self-inquiry method centred around the question, "Who am I?" This jivanmukta (liberated spirit[10]) spent his life in silence and contemplation, attracting over time, and despite his wishes, an increasing number of disciples. He would ask those who approached him, "Who is asking me the question?" If they answered with a name, he would inquire who was behind that name. If given a profession, he would ask who was behind that profession. Gradually, he invited his interlocutor to seek the answer within themselves, in pure consciousness (sat-chit-ânanda or Being-Consciousness-Bliss). In one of the few transcribed and translated interviews, he unveils the core of his doctrine:

> *I am not this physical body, [...] nor the five sensory perception organs, those being the ear, the eye, the tongue, the nose, and the skin, and their corresponding functions: hearing, sight, taste, smell, and touch. I am not the five organs of activity, that is, the vocal organs, hands, and feet, the procreation organ and the anus, and their respective functions: speech, body movement, enjoyment, and excretion. [...] Even the thinking mind, I am not; nor this state of ignorance where only object impressions exist, not the objects themselves and their functions. [...] After dismissing everything mentioned above as being "neither this nor that," the pure consciousness that remains—THAT I am*[11].

It is crucial to differentiate the I from living consciousness in general. Though these terms are often used synonymously, consciousness is merely the fact that something mental exists rather than nothing—whether I sense it or not. It is a pure THAT, impersonal and timeless, which some call the Being, others God, still others the Universal Mind or Spirit. These terminologies appear equivalent, but saying "my consciousness" is, strictly speaking, a convenient linguistic abuse to denote the I.

An illuminating analogy would compare consciousness to geometric space, containing all possible material phenomena without reducing itself to them. Though some individuals, after enclosing a plot of land, believe themselves proprietors of the encompassed area, they are not—and thank goodness!—proprietors of Space with a capital S.

While consciousness is universal, nature dictates that we traverse only this small enclosed portion of conscious experiences constituting the I. However, the limits of the I are not evident in our experience, much like the Universe's visible boundaries. I can imagine there being an instance originating all my singular experiences, yet I cannot apprehend this unity in its entirety. Sartre summarised this idea aptly in *The Transcendence of the*

Ego, expressing in essence that consciousness is something lived, not something contemplated. That's why Ramana Maharshi rightly states, "THAT I am," not "I am THAT."

All the fences I perceive encircling my conscious experience are necessarily limiting. The things the I identifies with at a given moment are the Me. If I therefore am the act of sensing something rather than nothing, the Me, on the other hand, comprises the stable sensations to which I identify. I would even make a distinction between the worldly Me, which is what I identify with as a person of the world; and the metaphysical Me, which is what I might hope to identify with after subjecting the worldly Me concept to philosophical critique. Let's clarify this thought.

The I can be seen as a window onto the world, an aperture through which a sensory content presents itself for contemplation. Metaphorically, this window constantly moves. The unfolding landscape before it is never the same. For if it were static, this landscape would preclude any possibility of experiencing being an I distinct from what is seen. I would be at most a part of the landscape. And so it goes with all things.

When I contemplate a chair, at that precise instant, I am nothing but the impressions of the chair: the view of its backrest, the scent of pine, the feel of my hand on the armrest, and the bamboo legs—all these form my conscious experiences. Perhaps this is why Huxley declared: "I spent several minutes—or was it centuries?—not merely contemplating these bamboo legs, but actually being them," writing this under the influence of mescaline, a drug ritually employed in Amerindian shamanism. He further adds: "or rather being myself in them; or, to be even more precise (for the "I" wasn't part of this affair, nor, in a certain sense, were they "they" themselves) being my not-self in the not-self of that armchair[12]."

From the perspective of lived experience, the world imposes itself on the I in its monolithic unity. I cannot immediately distinguish my hand from the chair, no more than I can separate the armrest from the rest of the perception. It is only because, in the following instant, I contemplate something else that the possibility arises of not being solely this chair I see; but indeed the very principle of contemplation underneath.

Let us not fall from Scylla to Charybdis: though the I's content varies over time, it cannot alter entirely. If each of the world's impressions at any given moment bore no relation to previous moments, it would equate to becoming a different being with every new perception. Here the scenario resembles a thought experiment proposed by Leibniz: in his *Essays on Theodicy*, he asks the reader if they would agree to relinquish their memory and be crowned king immediately after. A worldly individual might answer affirmatively. But considering the situation clearly, wouldn't it amount to dying and witnessing a different person, bearing the same name and appearance, ascend to the throne? Ensuring that the I can become self-aware necessitates that consciousness contents change over time while establish-

ing continuity between worldly sensations at each moment. "The *I think* signifies: I link," said Martin Heidegger[13].

The fixed points of sensation

We deduce that between two world impressions A and B, there must exist at least one common element and one distinct element. Without this condition, A and B would result either in a single static moment, lacking any distinction between the I and the world; or in two unrelated worlds, contemplated by two separate I's. These two I's drown, each on its own, in the contemplation of their respective objects, erasing all distinction between the subject and the perceived world.

Furthermore, it is because I only exist in the continuous change from A to B that when I think at any given moment, the moment has already passed. The present's intelligibility is contained in the immediate future. Thus, the instant is an abstraction of thought without which time couldn't be conceptualised; and when one unfortunately attempts to think this instant, it loses its meaning. This prompted Saint Augustine to say: "If no one asks me what time is, I know; if sought for explanation, I know not[14]."

From this discussion, we conclude there must be *fixed points* of sensation. Not fixed in the sense that they're perpetually present at all times, but because they ensure the link between moments pairwise. These fixed points are the I's anchors, its metaphysical moorings. They aren't always present, but some can become conscious when attention, through various exercises, is focused on them. They accompany the I continuously.

In lived experience, the most obvious fixed points are bodily sensations: the legs ensuring spatial movement, the senses allowing perception, the hands grasping objects; kinaesthesia and proprioception, sensing our limbs, their relative positions, and the internal actions of organs. It's the blood flowing through our veins, pulsing in our temples, beating slowly or like the frenzied drum of the depths; the breath whispering gently from the belly to the nose's tip, not forgetting the sensation of skin in nocturnal gales, the earth's pressure against our feet, flesh burning under the sun's caress, head spinning on its axis, clothes upon our back, elbows—the neck! This royal procession of sensations, ever-present with the I, ensuring dominion over the kingdom of living consciousness, are intrinsic parts of the Me.

These fixed points, however, are not world objects but privileged connexions between sensations. For if I consider them as contents of consciousness, the body, head, neck, hand sensations, legs—all lack subjectivity. They are complete objects of perception, indistinguishable, as such, from table or snuffbox. They are lived realities, manifestations of consciousness, not the entire I.

The body's essence isn't its form, colour, or spatial occupation. The body ensures a connexion between sensations. When my left hand touches

my right, as envisioned by Merleau-Ponty, I experience feeling one through the other. This magic, peculiar phenomenon doesn't occur when my hand strokes a creature's fur. In this case, I feel my hand on my cat's coarse, damp, or rough fur, but I don't feel what the fur senses upon my hand's touch. The fixed point is thus the sensation itself accompanying each moment of consciousness. More precisely, a fixed point is the privileged connexion between different sensations.

To illustrate this point, one might ponder what transpires if my hand became numb? If my eyes ceased seeing? Shoulders, arms, legs, sex lost their ability to feel? If the entire body became a shapeless rock, a bitter, savourless muck, incapable of bearing the world within like fresh pastry dough ready for the oven? Pushing this thought experiment to its limits renders a state equivalent to subject extinction. It would be a theoretical death so brutal and enigmatic it would be near-impossible to pinpoint where it begins. As long as a single thinking cell remains, a centimetre of sentient skin, a hair atop the skull's pinchable surface, the subject maintains the I. But once the final fixed point vanishes, it ceases.

If fixed points were parts of a whole, akin to cells forming organs or organs comprising a body, or thoughts forming parts of consciousness, one could mentally subtract them one by one. One might at least contemplate their subtraction. However, proceeding this way yields the absurd conclusion that removing something from the sensation whole causes sensation itself to vanish entirely. It follows that either (1) fixed points embody the entire whole or (2) they individually manifest within the whole. In the first case, fixed points being the I, for the I indeed denotes thinking, sensing, being something rather than nothing, indicates sensation, all conscious acts *possible*. Claiming fixed points as the whole renders the fixed point concept redundant. In this light, I might have been better off never introducing it!

It must thus be that fixed points manifest within the whole without being the entire whole. How is this possible? At times, philosophy's puzzles find answers in life's simplest pleasures. If a chef shows you an appetising cake, claiming a secret ingredient is within and adds, to pique your curiosity, that this ingredient appears in no specific cake piece but in each slice, what will you say? That it is the salt and life's bond. I propose seeing fixed points like milk, sugar, and eggs, finely blended, filling the world's full volume. They are connexions existing between all consciousness contents. By "connexion," I mean privileged associations. Each thought leads a retinue of other thoughts, images, sounds, scents, not through some divine dice throw but following habits' well-trod paths.

To conclude, I insist we often form an image of the Me, which I call the worldly Me, as divorced from the metaphysical Me as the moon from the Sun. The metaphysical Me doesn't exist as representation, but only as representation association. It is the story of the I. Not an objective story, woven of words, events. Not the library book tactile to hand, opened and

consulted. The metaphysical Me is the filter associating ideas based on past experiences. Hence, it acts as the prism through which the world opens to contemplation. Its profound nature is a conditioning. Fixed points are neither within the body, brain, nor sensation, but in the time that flows, like a river, from sensation to sensation; repeating patterns. On this account, perhaps one might say human existence balances delicate change with eternal return?

If it's true the body nears a fixed point more than any world object—it certainly does more than this book or cake before me—the body itself isn't a fixed point. Certainly, it accompanies each conscious experience. I can't feel, think, enjoy, or suffer except with presence of this body; so much so that seeking to remove it amounts to chasing one's tail. Yet what designates the body as the Me isn't the body as worldly object. It's the sensation link existing. The metaphysical Me isn't this worldly idea we gaze upon in mirrors. The metaphysical Me is immaterial, invisible, and indivisible. Thus searching for it in a body part, as Descartes or Swedenborg persisted in extracting the soul from the pineal gland, would be misguided. The idea of an objective Me is a mirage, a dream, an empty vision. The opium of philosophy. It prompts Ernst Mach to say:

> "The habit of treating the unanalyzed complex of the Me as an indivisible entity has often found curious scientific expression. First, one extracts the nervous system from the body, as the sensory seat. Then within this system, one privileges the brain, seeking in it a fixed point assigning the soul itself a seat to salvage the so-called psychic unity."

16.2.2 The Self and the Unconscious

Probably no one, confronted with the question "Who am I?" would respond with "I am the fixed points of consciousness," "I am the conditioning of my experiences," or even "I am the time flowing from one empty vessel to another." And even so, enumerating these fixed points is no easy task, for aside from bodily sensations (indeed the basics of meditation practice), one isn't necessarily conscious of fixed points. Although foundational to the Me and deduced *a priori*, they aren't always concrete thought objects.

Certainly, we form ideas of our body and its function, but such ideas poorly coincide with the body's entire complexity. Worse still: within the Me are elements surpassing visible, organic bodies. Psychological traumas endured during development, for instance, leave stigmas in the Me, manifesting as anxieties, somatic pains, or fantasies. These singularities of our inner world constitute our secret garden. It is on this fertile soil that characteristic reactions sprout, wherein we recognise our personality. Thus, in the visible aquarium of the Me, the traces of abyssal monsters, antediluvian

predators, ghostly eels glide, whispering discreetly over the bare wave—and it's the art of psychoanalysis to discern this invisibility.

When I, as consciousness, focus specific attention on my body and personality and apply concepts at my disposal to understand them, I forge an approximate Me idea. This occurs, for instance, when answering personality questionnaires and exclaiming with the fervour as if Archimedes had rediscovered flotation laws: "I am ENFP!" or "I have a type 9 enneagram!" It's also what unfolds following a social challenge, like a contest, interview, or imminent romantic encounter, demonstrating my competence, intelligence, beauty, sensitivity, frustration, or desirability.

Indeed, since the Me isn't directly visible, doubt often arises regarding my qualities, leading to comparisons against an imaginary norm. I seek, within psychotherapy or competition, validation of characteristics subtly suspected yet doubted deep down—with an odious impostor feeling in my heart. Relying on the other's gaze, I construct and verify a certain self-image. This self-image is what I shall call the Self. This image forms through definition, obtained with words, symbols, representations—when stating, "I am such a person, aged such, performing such a role."

Curiously, the first action associated with a newborn child, whose eyes, squinting from delivery, hesitate to escape primordial, limbo-like confusion, to unfetter themselves from that peace of beginnings; it's surreal to think of nothing else but assigning a name. Do we imagine these syllables conjure the subject's metaphysical condition? Believe that a syllable referring to the Mother of God or an apostle can save the Me from the invisibility to which it seems damned?

"The Me, in any case, cannot be saved," said Ernst Mach, but can it be named? The Self is a spell aimed at providing an objective, abstract, definitional, pragmatic unity to a saviour-Me, a rescuer-Me, who long surrendered appearing in the conscious world. Unlike the Me, the Self depends on myriad social games. It seeks Self-definition through interacting with others, particularly with significant people institutionally—teachers, judges, moral figures—or emotionally—partners, family members. The Self doesn't merely define its objective existence with a name, body, title, personality, history; it also seeks definition through ideas.

The Self clings, like inalienable appendages, to religious, political, philosophical opinions, tastes, aversions, and all kinds of limiting filters providing socially validated narratives. Thus, exchanging personal opinions or conceptions becomes painful, sometimes dangerous, where conversation seemingly benignly places interlocutors' identities at stake. Defensive reactions often aim to maintain the fragile construction the Self managed to achieve for certitude of being a tangible, real-world Me—and not the flow of disorganised sensations of a schizophrenic I. The Self is thus sometimes expandable, sometimes retractable, finding definition through group construction or opposing groups. Usually, these happen concurrently.

The remaining thesis will endeavour to show, step by step, how living consciousness forms a notion of the Self and why a part remains irreducibly unconscious. This requires, firstly, understanding how the I can objectify the sensory contents it traverses so these contents become determined perceptions. It entails theorising how, and under what conditions, this perception becomes a concept, then applying conclusions to perception's fixed points. For now, suffice it to grasp that the I is a dynamic producing, at times, an intimately, somatically determined largely unconscious feeling of the Me; at other times, a consciously constructed, socially, and intellectually conditioned Self.

We manipulate Me and Self concepts, daily like Monsieur Jourdain with prose: unwittingly. Conversely, the I demands philosophical thought exercise—or mystical, contemplative experiences—not immediately accessible. This difficulty likely arises since, at birth, the I isn't yet an I capable of conceptualising sensation's fixed points, establishing a stable concept, so the young human being lives in pronounced fusion with the world, within which they are only what they see, sense, feel. Yet a neuropsychological development event unfolds strangely. The I, inherently primary, providing any conscious reality with a foundation, fades in favour of the Me. The young child not only forgets that the world exists solely within them, but even when they remember—which inevitably occurs—they attribute consciousness' reality to the ever-present perception's fixed points under their gaze. They don't say like Descartes, "I think, I exist," rather they ponder, "Could the world be an illusion of *my* brain?" In doing so, they envision the brain as an objectivity of the world responsible for consciousness' subjectivity.

As is often in philosophy, what is primary logically isn't always chronologically so. The I is conceptually primary since it constitutes what allows no mental world, no form of Me; yet this I manifests only through consciousness contents obscuring its trace. Thus, the fundamental metaphysical stage where young beings become alternately sceptical and Pyrrhonian, denying the world's reality and their perceptions in turn, is also where living consciousness yields to a hybrid consciousness: a Schrödinger consciousness neither dead nor alive, neither objective nor subjective, neither spatial nor entirely mental. This metaphysical chimera is likely responsible for assertions like, "the mind is in the brain," "the soul is in the body," or in another vein, "the world only exists in my head."

Though initially captivating, and necessary for consciousness' hearty revival, such metaphysical assertions, e.g., "the mind is in the brain," constitute superficial real probes, incomplete I sketches. They tumble into the metaphysical chasms I outlined in section 12.3.2. Notably, they fail to acknowledge the fact that viewing the world as genuinely subjective renders it mere pure, non-material consciousness experience (existing "outside" or "parallel" to the body, brain, and objective space-time).

This purely subjective consciousness experience, shedding illusory objec-

tivity, faces an irreducible objectivity nonetheless: that there is something rather than nothing. Why-I? Where do I come from? Such are enduring questions. Thus, recognising the I doesn't free us from the objective world, nor elucidates all philosophical queries, nor even some simple questions, like why the Me dies when the head is severed from the body! The *real*, whether of atomic physicality or General Spirit, continues to be suggested as both the I's aim (I think the world) and its precondition (I arise from the world).

Despite significant metaphysical I neglect in adulthood, particularly in societies emphasized by personality externalisation (viewing the Me as worldly), humans retain, inscribed in their flesh with indelible ink, a vague, uncertain notion of a primordial truth, a being or essence, within conscious phenomena (acting behind and ostensibly originating them). This conviction, firmly rejected intellectually, persistently resurfaces in our hearts—especially when confronted with death or the fact that the Me holds no permanence.

Perhaps this is why, when a child is informed of their birth from parental wombs and eventual death, full satisfaction evades: "But what was I before that?" they ask, with eyes widened in self-evident wonder, awkwardly confronting parents, "and what will I be after?" The I's birth and death, whether real or not (an indeterminable), defy intellectual conception. For thinking I cease thinking, or feeling I cease feeling, remains a thought or feeling. It is the I's nature to *be* at all moments, simply, without future or past. Like being in Parmenides' poem:

> *But how what is*
> *could it be, later? How could it have become?*
> *If it became, it is not; no more than if it must one day be.*
> *Thus vanish unexplained genesis and death*[15].

Could this fusion conviction, albeit lost since childhood, underpin beliefs in life after death? Or pre-birth existence? Human consciousness, ever living in agony, remains an I seeking consciousness content for nourishment. When, by circumstances needing clarification, firmly established Me and Self barricades temporarily fall, the I's primary totality condition becomes visible. These limit experiences—be it passionate love, contemplation, psychoses, or substance absorption—permit seeing beyond the Me and retracing upstream to the mental origin.

Key Takeaways from this Chapter

Though Tononi's theory attempts to analogise the mind and matter's operational principles, it suffers from four weaknesses:

1. Introspection is immediate.
2. Consciousness is both unitary and composite.
3. Consciousness can be described statically.
4. Consciousness merely processes environmental information.

To overcome these impasses, I advocate for a mereological cognitivism. This new approach examines how objects form within the human mind; in this sense, it is a cognitivism, but the real viewpoint remains ontologically neutral.

What I mean is that questions aren't posed assuming external environment acts on the mind. The inquiry is understanding what internal necessity dictates a consciousness-formed distinction between objectivity (the material world) and subjectivity (lived experience).

This approach implies consciousness's essential property is being conscious of changing things. I term this principle differentiation. Furthermore, I develop a methodology to verify this principle's existence, consisting of studying the brain's isolation from stimuli (Cunningham's experiment).

From this methodology, I briefly outline how consciousness births a distinction between I, Me, and Self. I also describe laws governing psychic evolution, which will be expanded upon in my essay *Is Consciousness Dead?*

Conclusion

Do you think robots have a soul?

Philip K. Dick

Can a machine think? It is time to examine what the conceptual tools adopted in previous parts may teach us concerning the question at the core center of this dissertation. I have endeavoured — with no small diligence — to justify the existence of three strata in human thought: the level of *percepts* (sensation), the level of *intuitive concepts* (perception), and the level of *rational concepts* (abstract reasoning).

Abstract or rational concepts are defined solely through words or numbers. They are pure syntactic signs. They do not always correspond to any tangible or lived reality, yet they may be manipulated by means of logical reasoning. This level of thought most closely resembles operations of computers — and yet, paradoxically, it is the most superficial. When discussing Gödel's incompleteness theorems, I pointed out that, in any axiomatic theory of even moderate complexity, most propositions are syntactically undecidable: their truth or falsity cannot be determined from the axioms alone. TThe rare truths that logic does unveil are but isolated islets — lonely specks scattered across an infinite ocean. In other words, calculative reason is the skin of the fruit, the bark that veils the sap!

The level of percepts, by contrast, belongs to sensitive world. Perception emerges as a whole: a dense neighbourhood of events and spatial forms interwoven. There are no atomic units of perception — no pixels, no instants — that could serve as fixed coordinates. In the absence of such elementary bricks, the mind must isolate contrasts within the perceptual field, following guided by a principle of differentiation and assimilation among its parts. Forms arise — not always consciously. At most they are fleeting sensations, sometimes intense, but intimately tied to the context of experience, wherein it becomes difficult to distinguish the subject who feels from what is felt. One cannot isolate the percept and claim it exists outside the sensible experience itself. This layer of judgement is what neural networks attempt to emulate. Statistical correlations are drawn so as to give rise to categories within sensory data. Though the results may be impressive, I have endeavoured to

show how these algorithms fundamentally diverge from human perception in structuring the data.

The third level, situated between the other two, is that of intuitive concepts. Here, the human mind isolates a perceptual form and grants it an intelligible, conscious, and linguistically expressible reality. In this process emerges a distinction between the concept and the thing it refers to. One may adopt a Platonic standpoint, in which case the concept exists ideally and sensible forms are but its particular manifestations. But one may one might side with the empiricist, for whom the sensible precedes, and concepts merely synthesise it. Either way, the concept is a reality manipulable by thought, but bound to sensible content. It is neither a pure syntactic abstraction (as in Rationality), nor a mere singular sensation (as in Sensation).

I have suggested that the intuitive concepts of my theory occupy, in relation to Kantian thought, a position analogous to synthetic *a priori* judgements such as "The sum of 5 and 7 is 12" or "The sum of the angles of a triangle is 180 degrees." These judgements are grounded neither in pure reason nor in pure experience, but result from the way human beings structure phenomena. Kant's great insight, to phrase it with all elegance it deserves, is that sensible data are not perceived as they exist *in themselves*, but are shaped accordingly to our prior concepts and intuitions. These latter belong to the innate endowment of human being. When we think the universe is a Euclidean space or time is an objective reality, we behave like a overzealous engineer : having imprinted our tools upon all things, we believe Nature fashioned the world with those very same tools.

If the machine does not think as humans do, it is because it does not structure sensible phenomena with the same tools. To perceive this clearly, let us first observe how things appear within human perception, and then compare this to our algorithms. Within our mind — as I believe I have stated more than once — there exists an intimate relation between the whole of space and its parts. Points in space are not raw data; on the contrary, they draw their intelligibility from their relation to neighbouring points. There is thus an entanglement between space as a whole and the figures that unfold within it. This is what led Kant to claim that space is the receptacle of phenomena, that without it no body could be thought.

A similar, though more elusive, phenomenon occurs in our experience of time. Events are not given upon a chronological line extending from past to future. Human brain — which we attribute many a magical trick — does not merely line up stimuli in a queue. In the very experience of human mind, there are no instants *per se*, but only transitions unfolding across time intervals. This is why subjective time, for us, on the degree of change within our inner lives: when sceneries shift, when thoughts drift, when emotions run high — time seems to pass more quickly than when we remain in perfect stillness.

By contrast, algorithms handle sensible data in a radically different fashion. To begin with, they possess an internal clock — a mechanism that allows them to perform exactly one numerical operation per unit of time. Moreover, the machine cannot conceive of the relationship between the whole and its parts as interdependent. It slices information into atomic units of perception: an image is broken down into pixels, an audio recording is digitised, words are split into letters and read one by one (whereas human grasp the word in its entirety[14]). The machine's ontology is a *flatland* — a world with no relief — where knowledge is acquired by stacking elementary bricks. To put it in a lapidary phrase: the structures of human perception are holistic, while those of the machine are discrete.

Perspectivist Mysterianism

In analysing the problems posed by artificial intelligence—particularly those concerning judgement and intuition—I have not confined myself to a Kantian lens. On the contrary, I have critiqued Kant's notions of time and space as a priori intuitions and proposed in their place a new structural pair: the intuition of the Infinite and the intuition of the Whole. I then sought to formally characterise these intuitions through three mathematical functions:

1. The completion of infinite sequences.
2. The resolution of undecidable problems.
3. Luck when facing an NP-complete problem.

A classical Turing machine, comparable to the algorithms we use today, fails to execute these functions in any satisfactory way. However, I have shown that if we allow such a machine to count the infinite—by performing an infinite number of operations according to specific rules—then these three functions become realisable. This line of development led me to show how, precisely, the notion of infinity underlies the resolution of undecidable problems, and how the capacity to apprehend totality underlies NP-complete problems.

Taking a step back from the technical results, one could say that the intuition of the Whole and the intuition of the Infinite cannot be adequately modelled by a classical Turing machine. What is equally striking, however, is that it seems just as implausible to model the human mind using a transfinite

[14] A study from the uinervsity of Cbirmadge dmonstaretd that it is psoisble to raed a txet in wihch the ltetres are sracbmeld, so lnog as the fsrit and lsat ltteers of ecah wrod rmeian in the rgiht pclae. This dmeonsatrtes that the brain does not read all the ltetres but pecreives the wrod as a wlohe. Proof: admit you had no trouble reading this sntnece. (For a commentary on this paragraph, see Thierry Opillard, *La Dégradation Cambridge: Même pas vrai?*, Les Actes de Lecture no.102, June 2008.)

Turing machine. The reason lies in the fact that our brain, being finite in size, appears ill-suited to performing unbounded computational operations.

Everything thus unfolds as if there existed an irreducible mystery surrounding the human mind: the theories we possess for describing the brain fall short of capturing certain essential properties of intuition. Conversely, the idealised models we construct of the mind cannot be translated into physico-chemical terms. I have named this insight *perspectivist mysterianism.*

Within perspectivist mysterianism, we may entertain two types of representations concerning the mind. In one, the mind is intelligible, but the way it arises from matter remains obscure. In the other, the material functioning of the brain is comprehensible, but not the way in which it gives rise to certain specific types of cognitive properties. I have discussed this hypothesis in light of recent developments in neuroscience, and found in the works of Edelman and Tononi credible support for my philosophical vision.

Strong and Weak Thesis

My dissertation leads to two possible thesis about machine judgement — one weak, the other strong. The weak statement holds that classical Turing machines do not function like human mind; the strong one asserts that they're incapable of even imitating us. Let's look at both.

The weak thesis points to a difference in *functioning*, grounded in what I've called informational ontology — that is, in the nature of information: how it's structured, represented, and interconnected. In short, what it *means* to be a machine. Despite its name, this "weak" thesis has significant implications. It challenges the assumption — often implicit — that the mind is the brain, and that the brain is a kind of neural network. It invites us to reassess our computational reflex, the way we tend to model human cognition based on our favourite tools (computers, software...).

Thèse 3 (Weak Thesis)**.** *The human mind does not function like existing algorithms. Artificial neural networks are not accurate models of the human brain ; and no model of human brain is an accurate model of our mind.*

The strong thesis goes further: not only is the mind's ontology distinct from that of the machine, but its operations cannot even be imitated. The mind is a black box — not just in practice, but in principle — not because we lack the tools to peer into it, but because it operates in a fundamentally different way. All we can do is observe the correlations between its inputs and outputs. But this, precisely, is what neural networks do: they replicate a function that is either (a) unknown or (b) non-computable.

This doesn't necessarily mean that some mental tasks are forever inimitable. Rather, the strong thesis claims that, for any given machine, there

exists a task where it fails to reproduce human cognition. This more modest formulation draws from Gödel's incompleteness theorems (chapter 5.2) and Rice's theorem (5.3): not that some mathematical truths are unknowable or that some functions are uncomputable in an absolute sense, but rather that no single system — no universal Turing machine — can decide all possible problems. The dream of building a machine that could fully replicate human judgement — a "superintelligence" capable of grasping the totality of human knowledge — is precisely what the strong thesis denies.

This idea aligns closely with the artificial IQ tests devised by François Chollet, as discussed in chapter .

Thèse 4 (Strong Thesis)**.** *The human mind cannot be imitated by existing algorithms. For any algorithmic model, there exists a task where it fails to replicate human judgement.*

What I've tried to show is that the weak thesis, even if not definitively proven, raises enough problems to disqualify reductionist views that equate thought with computation. My analysis of human intuition — drawn from optical illusions and neuroscientific experiments like chronostasis — calls for a response from proponents of strong AI. And I hope they'll respond. Even if I'm mistaken, perhaps a serious debate can bring us closer to the truth. If needed, I'll step into the philosophical arena, holding up a sign that reads: *prove me wrong.*

Whereas the weak thesis is grounded in philosophical reasoning, the strong thesis belongs more to the realm of empirical research. It's impossible to determine *a priori* how well our statistical methods can imitate a function that, by its nature, we don't fully understand. And that's precisely the issue: because the function is unknown, we can't know whether it will turn out to be easy or difficult to reproduce. Even if we had a statistical model that perfectly mimicked the mind's behaviour on all known cases, we'd still have no assurance it would perform as well on unknown ones. This is the problem of *extrapolation*, and it can be illustrated with a simple example.

Let's suppose we're training a neural network to model the temperature at which a material becomes gaseous, depending on pressure. For simplicity, let's say the network has only one layer and performs linear regression — it looks for a linear correlation between input and output. In reality, the relationship between pressure and temperature follows a linear trend up to a certain point — the triple point, where solid, liquid, and gas coexist — but then the slope changes (see figure 16.2).

Our neural network will most likely fit a straight line approximating the average slope of the data, but it won't anticipate the shift at the triple point. This is because regression doesn't reveal the underlying laws — it only fits to the observed data. A more sophisticated network might detect the slope change. Perhaps humans themselves are just doing advanced statistics when they theorise the world. But scientific thought also involves choosing

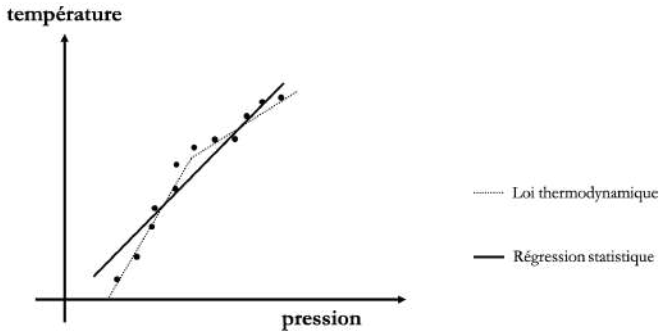


Figure 16.2: The temperature at which a substance becomes gaseous depends on pressure. Statistical regression fails to account for the change in slope predicted by thermodynamic laws.

which variables to use and designing meaningful experiments. If we feed the machine raw data with no structure or hypothesis, its predictions are likely to drift from reality. In this sense, a sharp critique of artificial intelligence ends up doubling as a sharp critique of science itself.

None of this rules out the possibility that machines might eventually imitate human cognition. But it does help identify the obstacles. If I had to bet — let's say with a gun to my head — I'm not entirely sure which side I'd choose. My feeling is that true imitation is, for now, beyond reach. Not because it's metaphysically impossible, but because: (1) we still don't understand how the mind works; (2) the machines we use are based on discrete logic; and (3) Gödel and Turing have shown us the limits of formal systems. But maybe that's just a hope, a belief — a cry in the dark.

If I've made the reader question some of today's strongest claims — that "AI is a thinking machine," that it's growing at an "exponential rate," or that "the singularity is near" — then this work will have achieved its aim. Before closing, I want to briefly sketch a few ethical questions raised by the increasing role of AI in human judgement.

ChatGPT: An Alchemy of Knowledge

Whether or not machines are capable of judgment should determine the very orientation of algorithmic design. Can it be reasonable to entrust decisions to programs incapable of understanding—except through statistical association—the intuitive concepts of justice, freedom, and dignity? Software such as ChatGPT appears capable of using these words to weave narratives. But do we judge a politician by their speeches alone? Or rather by the vision they offer for the society to come?

ChatGPT has stirred up such a storm—and its self-learning proceeds at

such an astonishing pace—that one question seems to burn on the lips of many in the digital world: might artificial intelligence pose a threat? The performances of chatbots, when tasked with giving a correct answer to a mathematical problem or offering a reasoned opinion in a political discussion, swing wildly. At times, the programme delivers accurate information; at others, it fabricates content—yet always with the same semblance of plausibility. This professional conjuror, who blows both hot and cold, cannot stand firm before an accredited expert, yet still astonishes with its uncanny ability never to capsize entirely.

Some might imagine that a machine's responses drift like wind and rain, with all the unpredictability we usually reserve for the weather. And yet, insofar as human beings themselves—modern humans, steeped in the wonders and woes of the internet—are no longer guaranteed bearers of truth; and insofar as our societies readily place impostor-scholars upon pedestals, one might ask whether AI is not, in the end, simply a human being: capable of error and genius alike, of precision as well as falsehood? It encapsulates, in a single figure, the George Psalmanazar of geography, the Grey Owl of Indigenous advocacy, the Charles Ponzi of finance, the Jean-Claude Romand of medicine—or more contemporaneously, the Stéphane Bourgoin of serial killer lore.

A cautious and nuanced appraisal of ChatGPT's abilities would distinguish its *technical* dimension from its properly *epistemological* one. From a technical standpoint, the achievement of its creators is without precedent, anticipating—perhaps by several years—the expected advances in the field of chatbots. One need only compare the performance of ChatGPT-2, which was racist, denied the Holocaust, and claimed Hitler a more competent statesman than Obama in fighting terrorism, with that of the far tamer ChatGPT-4, which is perhaps so well-behaved it now censors itself at the drop of a hat. Today's AI even passes Ben Goertzel's *Robot College Student Test*, which consists of impersonating a student in a university setting and successfully completing examinations. Although the machine is not physically present in classrooms, it performs admirably in assessments. It even passes, with flying colours, the entrance exam for engineering school[15]!

Despite this technical triumph, the epistemological record is more mixed. ChatGPT, far from being the revolutionary tool hailed in Dionysian hymns, might in this respect be merely the prolongation of existing search engines. Let us be clear. From the standpoint of a computer scientist, ChatGPT generates responses not only more quickly than a simple Google search, but often with greater accuracy—in answering, for instance, how to generate a particular snippet of code. For the twenty-first-century programmer, it is generally more efficient to consult ChatGPT than to phrase the same

[15] Axel Legay in https://www.rtbf.be/article/chatgpt-lintelligence-artificielle-qui-inquiete-les-profs-elle-reussit-lexamen-dentree-pour-les-ingenieurs-11122991

query on a search engine or to post it on a forum. Nonetheless, from a strictly conceptual point of view, there is no fundamental difference between these two operations. In both cases—setting aside the crucial questions of traceability and intellectual property—the goal is to retrieve an answer from the pool of data amassed across the Internet. The discourse of ChatGPT thus appears as a neatly packaged synthesis of the content provided by internet users—with all the accompanying errors or flashes of genius that were already there to be found.

This extension of the search engine thus exposes us, at an accelerated—not to say hyperbolic—rate, to both the promises and the perils of the internet. To mention only a few criticisms, there is a risk, for individual users, that the very process of seeking information online will, in the long run, engage less and less of the user's own capacity for judgment; that sources will become less traceable; that justifications will be fewer or harder to verify; that opinions will tend toward a more homogeneous consensus.

At the systemic level too, there is cause for concern that ChatGPT could be employed to influence public opinion. Unlike Wikipedia, which functions as a free platform aiming, through consensus, to extract objective truths, one can imagine an economic model based on the subtle diffusion of advertisements or certain ideologies. Conflicts could arise over making particular responses more visible through the chatbot, much as companies now jostle for position in Google search rankings. It is perhaps worth recalling, in this regard, that Google has developed an algorithm capable of recognising, with varying degrees of success, content generated by ChatGPT—with the aim of excluding from its indexing system pages whose content consists solely of automatically generated text.

This concern becomes critical if we dare to imagine what the internet might become once the overwhelming majority of pages indexed by search engines are nothing more than lines of ChatGPT output. Not only would it become nearly impossible to trace the origin of information, but worse still, algorithms dependent on big data would be left to sift through content generated by algorithms themselves—with all the ethical distortions this feedback loop could entail. It would be the equivalent of a mill attempting to produce flour using only its own grain!

Artificial Apocalypse or Last Judgement?

A final danger upon which I would like to dwell—this time of a political nature—would be that of witnessing an ever-higher normalisation of opinions and behaviours, or an ever-greater monopoly held over the vectors of information. It is well understood that the internet, far from broadening our minds through exposure to the diversity of sensibilities and beliefs which make up humanity, also operates to radicalise opinions, to confine us each

in belief spheres impervious to one another. Anyone who has spent time on the web will, at least once in their life, have encountered the staggering redundancy of information. To so innocent a question as: "what is a narcissistic person?", one finds a plethora of articles on Doctissimo, Aufeminin.com, and psychologies.com. Despite their purported diversity, all of them draw, directly or at a remove, upon the DSM-5 and ICD-6, the two principal manuals within psychiatric diagnostics. It sometimes even occurs, as one's digital wanderings meander through the net, that entire paragraphs are discovered copied-and-pasted from one article to the next.

This experience of the duplication and repetition of knowledge is, incidentally, mirrored in university study. Whatever field one chooses to pursue, one eventually realises—some years into academic apprenticeship—that the keystone works of a discipline tend, more or less unfailingly, to trace the same time-worn pathways of thought. Do you wish to understand what "attention" is, what constitutes "dialectic" or "intelligence"? In the state of the art, you shall discover an inclination to quote the same Socrates, Posner, or Binet, to draw the same distinctions and parallels, to endlessly voice the same critiques. Perhaps this is the very reason why, nowadays, it is so easy for a lecturer to correct two thousand exam scripts and supervise ten theses each year? Much as Descartes, with a mere glance at a table of contents, could divine what a book would discuss, so the person of the 21^{st} century may content themselves with nothing more than the title of an essay.

Yet, one might ask if this phenomenon—the normalisation of ideas, and consequently of the behaviours those ideas engender—arises more from the nature of the internet itself than from the statistical sum of the world's billions of individuals? Even supposing that collective intelligence naturally produces a convergence of intellectual pathways, it is difficult to deny that mass communication tools, crystallised in Google and Wikipedia—and without which such collective intelligence could not so easily have taken shape—induce a notable homogenisation.

It is not hard to imagine a future society in which the imitation of pictorial or textual works is rendered so effortless that original material productions, and the skills required for their creation, are held ever more precious. Rather than focusing innovation upon the elaboration of new ideas, and patenting inventions on the strength of the theoretical principles underlying their operation, it will be more vital to realise tangible products upon concrete supports. Possessing an artwork with one's own hands will, in this context, be deemed a luxury. The elaboration of certificates of authenticity, guaranteed by means such as NFTs (non-fungible tokens) or other digital processes, would become a matter of international legal and political import.

Alongside this initial tendency—and by extrapolating the trend—one might dream (or, conversely, grow wary) of a society in which ideas and intellectual works are no longer a matter of personal property, but form part of the common treasury of humanity. In such a world, it would be

presumptuous to lay claim to the paternity of an idea, just as it is arrogant, today, to lay claim to universal knowledge. What was once the very aim of intellectual labour would then pass for an eccentricity. There would no longer be such things as personal works, nor as individual labour.

Yet this civilisational universe, sketched in one broad stroke, stands in stark contrast to a doxa still steadfastly held in today's world—a point Hannah Arendt so aptly illustrated: we are (still and always) a society of labour[16]. And even if labour, asymptotically, passes into the domain of the machine, it remains from this labour, and its resultant produce, that the individual justifies their social and economic standing. Thus, one must either abandon the notion that labour constitutes personal value—to partially repeal the regime of intellectual property—or else discover a way to reinvent labour itself.

For it is crucial to remember that a technical instrument is never more than the tip of the iceberg. Behind the operational modes of an algorithm lurks also, often more insidiously, a certain vision of society. This "ruse of digital reason", if I might paraphrase Hegel, is responsible for the fact that, unbeknownst to its designer, software conveys social transformation regardless of intent. Thus it is that the internet has done more for Homo Sapiens than merely provide cyclopean databases. It is not but one more book upon the world's dust-laden shelf. Digital technologies—including ChatGPT—are, more precisely, reshaping the very distinction that once separated public from private space; opinion from behaviour; judgement from decision.

Within the Enlightenment paradigm, the individual was as master and possessor of their own intellectual intimacy. Indeed, it is in the modern era that the term "intimacy"—which in Medieval times referred solely to the affective bond with a third party—came to designate that which is personal, private[17]. Yet today, the algorithms of social networks, though they cannot utilise your data for all purposes, nonetheless know far more about your political opinions, your sexual orientation, your psychological disorders, your digital footprint, than your best friend might ever suspect. We no longer sell solely our labour power (let it be noted that the human effort required to satisfy vital needs diminishes ever more substantially)—we sell our desires, our dreams, our fears, as if we ourselves, in the deepest layers of our subjectivity, have become a commodity.

In this respect, legislations such as the GDPR are, more than instruments for the protection of personal data, confessions of impotence. They constitute the final convulsions of an institutional regime overtaken by the vertiginous pace of technological development. To believe that our data be-

[16] See the Prologue to *The Human Condition*.

[17] For a genealogy of the meanings of the word "intimacy" through encyclopaedias from the seventeenth century to the present, see Isabelle Berrebi-Hoffmann, *Les métamorphoses de l'intime*, in Empan 2010/1 (n°77), pp.13-17.

longs to us by virtue of consent is to conjure a kind of parallel reality. In the legal universe, which applies to individuals endowed with names, faces, and flesh-and-bone bodies, all proceeds as if intimacy underpins human dignity. But in the virtual universe—comprised of numbers, statistics, predictive models—our subjectivity has long since been stripped of all that made it personal. The contemporary world, more than ever, is schizophrenic. If "power" signifies both the descriptive capacity to influence others and the prescriptive capacity legally assigned by moral norms, then it is fair to say that "power" has been split into two antithetical entities: the real power held potentially over personal data far exceeds the legally permitted use. It is, indeed, by virtue of this separation—however theoretical—that this commercial regime of illegality, or all-legality, finds its efficacy and justification.

Be that as it may, ChatGPT seems to raise, today, a primary question: upon what principles do the powers of information synthesis deployed by artificial intelligence operate, and in what ways are these liable to influence the cognitive functioning of human beings? One has seen that, for Kant, judgement is precisely that faculty by which concrete impressions are brought into the objective unity of apperception. If we dispense with the technical, somewhat esoteric vocabulary, judgement thus refers to the capacity of the thinking subject to gather images, words, phrases, thoughts, or sensations under an encompassing perception. It cannot, therefore, be reduced to rational definitions or statistical associations. And yet, as things stand, there is every reason to doubt that AI is capable of producing subjective meaning out of an objectively linked sequence of words which it is so adept at aligning. Do we truly wish to entrust the governance of our societies and behaviours to a judge without eyes, without ears, without a soul? We might gaze upon such a judge with, upon our lips, the celebrated words of Albert Einstein:

> *In our efforts to understand [Artificial Intelligence], we are somewhat like a man trying to comprehend the mechanism of a closed watch. He sees the face and the moving hands, he even hears its ticking, but he has no means of opening the case*[18].

[18] I have taken the liberty of substituting the term "reality," as found in the original quotation, with "artificial intelligence" (which I have enclosed in brackets). See Albert Einstein, Leopold Infeld, *The Evolution of Physics*, 1938.

Bibliography

[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, Ladrange, Paris, 1866.

[8] Aristote, *On Ideas: Aristotle's Criticism of Plato's Theory of Forms* (ΠΕΡΙ ΙΔΕΩΟΝ), 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, April 2019.

[15] Barot Emmanuel, *En quoi la crise des fondements des mathématiques est-elle terminée ?*, Philosophia Scientiæ [En ligne], 9-2 | 2005, 15th June 2011, acceded the 17th of October 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, 7-9th October 2011, p. 46.

[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, 16th May 2020, https://intellectualmathematics.com/blog/first-proofs-thales-and-the-beginnings-of-geometry/ (acceded the 6th November 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.

[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, 5th November 2019.

[34] Chollet François, *ARC-AGI*, acceded the 17th of October 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] Cotnoir Aaron J., Varzi Achille, (2021) *Mereology*, Oxford, OUP., ISBN 978-0-19-874900-4, DOI 10.1093/OSO/9780198749004.001.0001.

[41] Crick F., *Astonishing Hypothesis: The Scientific Search for The Soul*, Cambridge University Press, Cambridge 1995.

[42] Damasio R. Antonio, *L'Erreur de Descartes*, Odile Jacob, 1997.

[43] Damasio R. Antonio, *Spinoza avait raison. Le cerveau de la tristesse, de la joie et des émotions*, Odile Jacob, 2003.

[44] 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.

[45] Dehaene et al., *The Mental Representation of Parity and Number Magnitude*, J. Exp. Psychol., 122, 371–396, 1993.

[46] Stanislas Dehaene, *L'accès à la conscience* Discours tenu au Collège de France en 2009-2010. Vidéo YouTube (acceded the 17th October 2024) : https://www.youtube.com/watch?v=Zt_fMVnbDYE.

[47] 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.

[48] Deleuze Luc, *Lettres et autres textes*, édition préparée par David Lapoujade, Paris, Les Éditions de Minuit, 2015.

[49] R. Descartes (1649) *Les Passions de l'âme*. France: n.p., 1728.

[50] Descartes, *Discours de la méthode*, annoté par Laurence Renault, Paris, Flammarion, coll. « GF », 2016.

[51] Dewalq Arnaud, *Deux théories de l'analyse psychique : Wundt et Brentano* in *Vers une philosophie scientifique : Le programme de Brentano*, Paris, Demopolis, 2014.

[52] Philippe K. Dick, *Les androïdes rêvent-ils de moutons électriques ?*, Éditions J'ai Lu, 16 février 2022.

[53] Dreyfus Hubert L., *Mind Over Machine*, Free Press, 12th September 1988.

[54] Dreyfus Hubert L., *What Computer Still Can't Do ?*, Harper & Row Publishers, New York, 1972.

[55] Earl B., *The Biological Function of Consciousness*, Front Psychol 5:1–18, 2014.

[56] Edelman Gerald M., *Antibody Structure and Molecular Immunology*, Nobel Lecture, December 12, 1972/

[57] *The Remembered Present : A Biological Theory of Consciousness*, 1989.

[58] Edelman Gerald, *Bright Air, Brilliant Fire : on the Matter of the Mind*, Penguin, Harmondsworth 1997, pp.109.

[59] Einstein Albert, Infeld Leopold, *The Evolution of Physics*, 1938.

[60] Edelman Gerald, Tononi Giulio, *Consciousness and Complexity*, December 1998, 282(5395):1846-51.

[61] Edelman Gerald , Tononi Giulio, *Comment la matière devient conscience ?*, Odile Jacob, 1st May 2000.

[62] Edelman Gerald, *Naturalizing Consciousness: A Theoretical Framework*, PNAS, April 2003, 100 (9) 5520-5524.

[63] Edelman Gerald, *Plus vaste est le ciel : Une nouvelle théorie générale du cerveau*, Odile Jacob, Paris, 2004.

[64] Edelman Gerald, *Second Nature: Brain Science and Human Knowledge*, Yale University Press, 2006.

[65] (1964) *Eliza*, program, `https://www.masswerk.at/elizabot/eliza.html` (acceded the 17th October 2024).

[66] 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.

[67] 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

[68] Evan Peter W., *A study of Time in Modern Physics*, Doctoral Dissertation, The University of Sydney, Philpaper, 2011.

[69] Fallon James, *The Psychopath Inside: A Neuroscientist's Personal Journey Into the Dark Side of the Brain*, Hardcover, October 31, 2013.

[70] Fingelkurts AA, Fingelkurts AA, *Information Flow in the Brain: Ordered Sequences of Metastable States*, Information 8(1):22, 2017.

[71] 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.

[72] Frege Gottlob, *Kleine Schriften*, I. Angelleli (ed.), Hidesheim: Georg Olms, 1967.

[73] Frege Gottlob, *Les Fondements de l'arithmétique*, trad. Claude Imbert, L'ordre philosophique, Seuil, 1969.

[74] 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).

[75] Freud Sigmund, *Esquisse d'une psychologie*, trad. de l'allemand, Toulouse, Erès Collection Scripta, 2011.

[76] Friedberg, R.M., *A Learning Machine: Part I*, IBM Journal of Research and Development, vol. 2, Issue 1, January 1958.

[77] Gardner Hendrie, *Oral History of Richard Greenblatt*, interview, 2005 Computer History Museum, CHM Reference number: X3131.2005.

[78] Gregory R.L. and Gombrich E.H., *The Confounded Eye* in eds. Illusion in Nature and Art, Duckworth, London, 1973.

[79] Gold I., Stoljar D., *A Neuron Doctrine in the Philosophy of Neuroscience*, Behavioral and Brain Sciences, 22(5), 809-30 ; discussion 831-69.

[80] Habermas Jürgen, Théorie de l'agir communicationnel, titre original : Theorie des kommunikativen Handelns, 2 t. (1981) ; réédition : Paris, Fayard, 1987.

[81] Haggard P, Clark S, Kalogeras J., *Voluntary action and conscious awareness*. Nat Neurosci. 2002;5:382–385.

[82] Babak Hassibi, David G. Stork, *Second Order Derivatives for Network Pruning: Optimal Brain Surgeon*, Ricoh California Research Center, Stanford, CA 94305.

[83] Georg Wilhelm Friedrich, *Phénoménologie de l'esprit de Hegel*, Poche, 2012.

[84] Heidegger Martin, *Être et temps*, trad. Emmanuel Martineau, Édition numérique hors commerce, http://t.m.p.free.fr/textes/Heidegger_etre_et_temps.pdf (acceded the 10th November 2024).

[85] Hodinott-Hill Iona, Thilo Kai V., Cowey Alan, Walsh Vincent, *Auditory Chronostasis: Hanging on the Telephone*, 15th October 2002.

Howdyshell Stanford,*On Realism and the Pessimistic Meta-Induction*, Kent State University, Critical Reflections, Essay on Significance and Critical Reflections.

[86] Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*, Basic Books, January 1980.

[87] Hume David, *An Enquiry concerning Human Understanding*, Londre, 1739-1740.

[88] John F. Ireland, Marvin W. Kahn, *How Fair Is the Culture I.Q. Test?*, Sage Journal, V.25-1, April 1979.

[89] James William, *The Principles of Psychology*, Holt, New York, 1890.

[90] Jaynes Julian, *The Origin of Consciousness in the Breakdown of the Bicameral Mind*, Houghton Mifflin, 1976.

[91] Jeshion Robin, *Frege's Notions of Self-Evidence*, Mind, Oxford University Press, Vol. 110, No. 440 (Oct., 2001), pp. 937-976.

[92] Johnson Aysja, Vong Wai Keen, Lake Brenden M. , & Gureckis Todd M., *Fast and Flexible : Human Program Induction in Abstract Reasoning Task*, arXiv, 10th March 2021.

[93] Jonkisz J., *Consciousness: Individuated Information in Action*, Front Psychol 6:1–17, 2015.

[94] Jonkisz J. *Subjectivity: a Case of Biological Individuation and an Adaptive Response to Informational Overflow*, Front Psychol 7:1–6, 2016.

[95] Jung Carl Gustav, Pauli Wolfgang, *Naturerklärung und Psyche*, Zurich Rascher, 1952.

[96] Kahneman (2011), *Système 1 et Système 2 : les deux vitesses de la pensée*, Macmillan. ISBN 978-1-4299-6935-2. Retrieved April 8, 2012.

[97] François Kammerer, *Conscience et matière*, Éditions matériologiques, Collection Sciences & philosophie, 2019.

[98] Kandel Eric R., *La Biologie moléculaire de la mémoire: un dialogue entre gènes et synapses*, M/S : médecine sciences, 19(5), 625–633.

[99] Kant Emmanuel, *Critique de la Raison pure*, trad. de Barni Jules, Édition Germer-Baillière, 1869.

[100] Kant Emmanuel, *Critique de la raison pure*, trad. Tissot Joseph, Librairie philosophique de Ladrange, édition originale 1836, édition révisée 1845 (Tome I) et 1869 (Tome II).

[101] Kant Emmanuel, *Prolégomènes à toute métaphysique future*, trad. Guillermit Louis, préf. Vuillemin Jules, Paris, Vrin, coll. « Bibliothèque des textes philosophiques », 2001.

[102] Kaye R., *The Minesweeper is NP-complete*,The Mathematical Intelligencer, Volume 22, number 2, 2000, pp. 9–1.5.

[103] 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.

[104] Khayyam Omar, *Commentary on the Difficulties Concerning the Postulates of Euclid's Elements*, 1077.

[105] 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.

[106] Jaegwon Kim, *La Survenance et l'esprit*, Broché – Grand livre, 9th January 2009.

[107] Kinzel Katherina, *Neo-Kantian Conceptualism: Between Scientific Experience and Everyday Perception*, British Journal for the History of Philosophy, 08 Aug 2023.

[108] Morris Kline, *Les Fondements des mathématiques*, *La Recherche*, March 1975, num. 54, vol. 6, pp. 200-208.

[109] Knuth Donald E., *The Art of Computer Programming: Fundamental Algorithms*, Addison-Wesley, 1997.

[110] Ladrière Jean, *Les Limitations internes des formalismes : Étude sur la signification du théoreme de Gödel et des théoremes apparentés dans la théorie des fondements des mathématiques*, Ed. Jacques Gabay, 1992.

[111] Yann Le Cun, Denker John S. et Solla Sara A. , *Optimal Brain Damage*, AT&T Bell Laboratories, Holmdel, N. J. 07733.

[112] Legay Axel, *ChatGPT, l'intelligence artificielle qui inquiète les profs : "Elle réussit l'examen d'entrée pour les ingénieurs !"*, https://www.rtbf.be/article/chatgpt-lintelligence-artificielle-qui-inquiete-les-profs-elle-reussit-lexamen-dentr (acceded the 5th November 2024).

[113] Lembezat Carole, *Google suspend un ingénieur qui assure que son IA est "consciente"*, Courrier international, 23th Jully 2022, https://www.courrierinternational.com/article/polemique-google-suspend-un-ingenieur-qui-assure-que-son-ia-est-consciente.

[114] Lepage François et Fradet Karine, *La crise des fondements : quelle crise ?*, Les Cahiers d'Ithaque, 7-9th October 2011.

[115] Leslie G. Valiant, *Knowledge Infusion: In Pursuit of Robustness in Artificial Intelligence*, Foundations of Software Technology and Theoretical Computer Science (Bangalore) 2008 (conférence), January 2008.

[116] Linsky Bernard, *Ernst Schroeder and Zermelo's Anticipation of Russell's Paradox* in *La crise des fondements : quelle crise ?*, Les Cahiers d'Ithaque, 7-9th October 2011., p. 46.

[117] Lobatchevsky Nikolai Ivanovich, *Nouveaux principes de géométrie avec une théorie complète des parallèles*, Kazan, 1835-1838.

[118] Locke, *Essai sur l'Entendement humain*, Traduction par Coste Pierre, Pierre Mortier, 1735 (p. Titre-321).

[119] 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).

[120] Lycan William G., *Consciousness*, First published October 1, 1987.

[121] 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.

[122] 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 (acceded the 5th November 2024).

[123] Marchetti Giorgio, *Consciousness: a unique way of processing information*, Cognitive Processing (2018) 19:435–464.

[124] Mansion Suzanne, *La Critique de la Théorie des Idées dans le ΠΕΡΙ ΙΔΕΩΝ*, Revue philosophique de Louvain, 1949.

[125] 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, 31th August 1955, https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (acceded the 17th October 2024).

[126] McCulloch, W.S., Pitts, W., *A Logical Calculus of the Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics 5, 115–133, 1943.

[127] McGinn Colin, *Can We Solve the Mind-Body problem?*, 1989, Mind, New Series, Volume 98, Issue 391 (Jul., 1989).

[128] Merleau-Ponty Maurice, *Phénoménologie de la perception*, Paris, Gallimard, coll. « Tel », 2005.

[129] Monier Cyril, *Les neurosciences au sein des sciences de la cognition : Une revue de la littérature plaidant pour un devenir spinoziste des neurosciences*, 2019.

[130] 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.

[131] Nabokov, *Lolita*, annotée par Alfred Appel Jr., New York, Vintage Books, 1991.

[132] Nagel Thomas, *What Is It Like to Be a Bat?*, The Philosophical Review, Vol. 83, No. 4 (Oct., 1974), pp. 435-450.

[133] A. Newell, J. C. Shaw, H. A.Simon, *Report on a General Problem-Solving rogram*, 30th December 1958, revised the 9th of February 1959.

[134] O'Neil Cathy, *Algorithmes : la bombe à retardement*, Broché – Grand livre, 7th November 2018.

[135] Opillard Thierry, *La Dégradation Cambridge : Même pas vrai ?*, Les Actes de Lecture n.102, June 2008.

[136] Ortigues Edmond, *Edelman et la conscience humaine*, Éditions de Minuit, dans *Critique* 2005-12 (Num. 703), 2005-12 (Num. 703), pages 917 à 926.

[137] Papadimitriou Christos H., Lewis Harry. R., *Elements of the Theory of Calculation*, Second Edition, Prentice-Hall, Upper Saddle River, New Jersey, 1998.

[138] Parménide, *Fragment du poème*, http://philoctetes.free.fr/uniparmenide.htm (acceded the 10th November 2024).

[139] Pascal Blaise, *Pensées*, L'édition électronique de Descotes D. et Proust G., https://www.penseesdepascal.fr/ (acceded the 17th October 2024).

[140] 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.)

[141] Peña Adolfo, *The Dreyfus Model of Clinical Problem-Solving Skills Acquisition: a Critical Perspective.* Med Educ Online, June 2010.

[142] 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.

[143] Georges Perec, *La vie, mode d'emploi*, Le Livre de Poche, 1st January 1980.

[144] *The Perl Design Patterns Book*, GNU FDL license, 1995.

[145] Perrin Michel, *Logique « sauvage » des psychotropes : le cas des sociétés chamaniques*, Psychotropes, 1991, VI, 4, pp. 115-123.

[146] Perrin Daniel, *Les Géométries non euclidiennes*, Conférence de Romilly, https://www.imo.universite-paris-saclay.fr/~daniel.perrin/Conferences/Romilly.pdf (acceded the 17th October 2024).

[147] 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.

[148] Platon, *Ménon*, trad. Cousin, Philosophie, September 2009.

[149] Platon *République*, trad. de Cousin, consultable sur *L'Antiquité grecque et latine du moyen âge*, https://remacle.org/ (acceded the 3th November 2024).

[150] Platon, *Théétète [ou sur la science ; genre peirastique]*, trad. Chambry E., La Bibliothèque électronique du Québec, Collection Philosophie, Volume 9 : version 1.01.

[151] Plutarque, *Dialogues pythiques*, trad. Ildefonse Frédérique, Éditions Flammarion, Paris, 2006.

[152] Poincaré Henry, *Les Mathématiques et la Logique*, Revue de Métaphysique et de morale, années 1905, pp. 815-835, 1906, pp. 17-38 et pp. 297- 317.

[153] Poincaré Henry, *La Science et l'Hypothèse*, Flammarion, Collection *Bibliothèque de philosophie scientifique*, Paris, 1917.

[154] Poirier P., Payette N., *Les Gardiens du bon usage*, Philosophiques, vol.34, n.1, pp. 183-200.

[155] Premack David, Woodruff Guy, *Does the Chimpanzee Have a Theory of Mind ?*, Behavioral and Brain Sciences, Vol. 1, N. 4, December 1978, pp. 515-526.

[156] Prinz Jesse J., *Level-Headed Mysterianism and Artificial Experience*, Journal of Consciousness Studies 10 (4-5):111-132, 2003.

[157] 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).

[158] Ptolémée Claude, *Almageste*.

[159] Quincey Thomas de, *Les derniers jours d'Emmanuel Kant,* éditions Mille et une nuits, 1996.

[160] Quine W. V. O., *Word and Object*, Cambridge, MA: MIT Press, 1960.

[161] Ramus Frank, *Un épouvantail bien commode*, Presses Universitaires de France, 2014.

[162] Reilly Douglas L., Cooper Leon N., Elbaum Charles, *A Neural Model for Category Learning*, Biological Cybernetics, 45, 35-41, 1982.

[163] 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).

[164] *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.

[165] *Rose, D., Summers, J. Duration Illusions in a Train of Visual Stimuli*, Scopus, Volume 24, Issue 10, 1995, Pages 1177-1187.

[166] Rosenfeld Boris, *A History of Non-Euclidean Geometry : Evolution of the Concept of a Geometric Space*, Springer, New York, 1988.

[167] 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.

[168] Roux Sylvain, *La recherche du Principe chez Platon, Aristote et Plotin*, Vrin, 2004, https://books.openedition.org/vrin/5432?lang=fr (acceded the 17th October 2024).

[169] Rumelhart David E., McClelland James L., *Parallel Distributed Processing. Exploration in the microstructure of cognition, vol I.* Cambridge, MIT Press.

[170] 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).

[171] 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).

[172] Russell Stuart et Norvig Peter, *Artificial Intelligence : A Modern Approach*, Pearson Education, troisième édition, 2003.

[173] Ryle G., *The Concept of Mind*, Hutchinson, London, 1949.

[174] Saertener Olivier, Guay Alexandre, *A New Look at Survenance. Or when After is Different*, European Journal of Philosophy of Science, 2016.

[175] 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.

[176] Sartre, J-P., *La Transcendance de l'Ego : Esquisse d'une description phénoménologique*, Librairie Philosophique J.Vrin, Paris, 1966.

[177] 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.

[178] 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.

[179] 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.

[180] Simon Herbert, *The New Science of Management Decision*, édition révisée, Englewood Cliffs, NJ: Prentice-Hall, 1977.

[181] Souchon Marguerite, *Le Dieu de Dostoïevski*, Éditions Première partie, 6th October 2021.

[182] 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.

[183] Taylor, *Paying attention to consciousness*, Progress in Neurobiology, 2008.

[184] Thomasson Dan, *Georges Perec's 10x10 Knight's Tour*, acceded the 17th October 2024, https://www.danthomasson.com/georges-perec.html.

[185] Tillemans T., *Dharmakirti* in Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*, Spring 2021 ed. Metaphysics Research Lab, Stanford University; 2021.

[186] Tononi G., *Integrated Information Theory of Consciousness: an Ipdated 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.

[187] Tononi Giulio, Boly Melanie, Massimini Marcello, Koch Christof, *Integrated Information Theory : From Consciousness to its Physical Substrate*, Nature Reviews Neurosciences, May 2016.

[188] Tononi Giulio et al., *Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms*, arXiv, 30th December 2022.

[189] Tononi Giulio et al., *Only What Exists Can Cause: An Intrinsic View on Free Will*, arXiv, June 2022.

[190] Tononi Giulio, *On being*, à paraître.

[191] Turing Alan (1950) *Computing Machinery and Intelligence*, Mind 49: 433-460.

[192] Turing Alan, *Intelligent Machinery*, Collected Works of A.M. Turing, Mechanical Intelligence, 1948.

[193] 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.

[194] Van der Linden Tim, *Notions de logique mathématique*, note de cours de l'UCLouvain, 19th May 2013.

[195] Van Gulick R., *What would count as explaining consciousness?* In T. Metzinger, ed. Conscious Experience. Paderborn: Ferdinand Schöningh, 1995.

[196] Varela Francisco, *La Recherche*, No. 308 April 1998.

[197] Verleysen Michel, François Blayot, *Les réseaux de neurones artificiels (Que sais-je ?)*, Presses Universitaires de France, 1996.

[198] Velmans Max, *Understanding Consciousness*, second edition, Routledge/Psychology Press, 2009.

[199] Vincent L., *Comment devient-on amoureux ?*, Paris, Odile Jacob, 2004

[200] Wernicke C., *Grundriss der Psychiatrie in klinischen Vorlesungen*, 1900.

[201] Wiener Norbert, *Cybernetics: Or Control and Communication in the Animal and the Machine*, Preface, page xi (footnote 1), 1948.

[202] 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.

[203] Yevarousky Valery, *The "Kantian Brain" Concept : Modeling the Limit of Human Perception and Existence*, September 2023.

[204] Zermelo, *Investigations in the Foundations of Set Theory I* (Abridged Version), Trad. Andreas Reif, 1908.

Appendix

Appendix 1: A Short Story Made by ChatGPT

The progress of ChatGPT has, understandably, been the talk of the town. Since its release in 2022, the AI has continuously improved, with an ever-increasing number of web pages being drafted using ChatGPT[19]. In this regard, the phenomenon may represent an inaugural existential threat to the supremacy and proper functioning of traditional search engines. To illustrate some of the strengths and limitations of ChatGPT, the reader will find below a short story entirely generated by this software (version o1 preview) as of 17 September 2024, preceded by the guidelines provided to the program.

This experiment was conducted in cooperation with Joseph Lange, co-founder of Fabuly, a start-up specialising in the creation of audiobooks that was organising a short story contest in September 2024. ChatGPT was given the contest guidelines[20]. The result was then translated into French using ChatGPT itself. The short story is titled "The Weaver of Whispers". Admittedly, there are evident naïvetés regarding the content, particularly its emotional aspect, but even the fiercest critics of artificial intelligence cannot deny the presence of a clear narrative structure.

Story Guidelines

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

[19] This phenomenon does not only affect web pages: a study published by Elsevier shows that, on average, 5.6% of the content in articles listed in November 2023 on *Orthopaedics & Traumatology: Surgery & Research (OTSR)*, a scientific platform aimed at medicine, had been generated by ChatGPT. This usage largely concerns "abstracts" (see *What is the rate of text generated by artificial intelligence over a year of publication in Orthopaedics & Traumatology: Surgery & Research?*)

[20] I myself wrote the contest guidelines, in exchange for which I requested to conduct some experiments with the paid version of ChatGPT used by Fabuly.

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 traveller 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 traveller, 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 Fabuly 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 changeable! 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 labours 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)

The Weaver of Whispers

The city hummed with its usual rhythm—blaring horns, distant sirens, and the murmur of countless conversations merging into one indistinct noise. Amelia navigated through the throngs on the pavement, her thoughts consumed by the monotonous routines which had come to define her existence. As a conservator of ancient textiles at the museum, her work suited her meticulous nature but left her yearning for something more.

One evening, while cataloguing a donation of items, she stumbled upon a singular piece—an ancient tapestry fragment, its edges frayed and colours faded. The design was intricate: a labyrinth of intertwined threads forming patterns that seemed to shift beneath her gaze. There was no record of its origin, no note of the donor. Intrigued, Amelia decided to examine it more closely.

Under the soft glow of her desk lamp, she noticed a barely visible symbol woven in a corner—a small stylised hourglass entwined with ivy. It sparked a childhood memory, a story her grandmother used to tell about a legendary weaver whose creations were said to possess mystical powers. Dismissing it as fanciful, she set the tapestry aside.

That night, sleep eluded her. Images of the tapestry danced behind her closed eyelids, calling her back. Yielding to the pull, she dressed and returned to the museum. The city streets were eerily quiet, a stark contrast to their daytime frenzy. The museum guard, used to her nocturnal visits, nodded as she signed in.

Back in the workshop, Amelia spread the tapestry across the table. As her fingers traced the fabric, a tingling sensation coursed through her body. The air thickened slightly, and the room seemed to emit a faint hum, though its source was indiscernible. Surprised, she stepped back.

"Who's there?" she whispered, her voice barely more than a breath.

Only silence replied. Shaking off her unease, she leaned in to inspect the tapestry again. This time, the patterns seemed different—subtle shifts in the weave she couldn't quite grasp. Determined to document her observations, she reached for her sketchbook.

Hours passed as she meticulously drew each detail. When she finally glanced up, dawn was breaking, light filtering through the blinds. Exhausted but satisfied, she gathered her things to leave. As she turned off the lights, a glow caught her eye. The tapestry seemed to emit a faint luminescence, but on closer inspection, it appeared ordinary once more.

In the days following, Amelia found herself preoccupied with the tapestry. She began to research the hourglass symbol, delving into archives and obscure texts. One afternoon, nestled in a dusty corner of the library, she stumbled upon a reference in an old tome on forgotten artisans.

“The Weaver of Whispers,” she read aloud, “a mythical figure whose tapestries were said to connect realms, linking past, present, and future.” The accompanying illustration bore the same hourglass entwined with ivy. A shiver ran down her spine.

That evening, as she pondered the implications, her phone buzzed with a message from her friend Leo, a historian with a passion for the esoteric.

“I've found something that might interest you,” read the message, followed by an address.

Intrigued, she agreed to meet him. The address led her to an antique shop she hadn't noticed before, tucked between a bakery and a bookstore. The bell above the door chimed softly as she entered.

“Amelia, over here,” called Leo from the back of the shop. He stood beside a series of textiles hung on the wall. “Look familiar?”

Her eyes widened as she recognised the patterns—similar to those of the tapestry she'd been studying. “Where did you find these?”

"The owner inherited them from a relative. I thought you might want to take a look."

An elderly woman emerged from behind a curtain. Her eyes were a striking green, sparkling with an energy that belied her age.

"You must be Amelia," she said warmly. "I've heard much about your work."

Surprised, Amelia replied, "I didn't know my reputation extended beyond the museum."

The woman smiled enigmatically. "Some things are meant to be discovered."

Amelia examined the textiles, noting the recurring motifs. "These designs—they're linked to a legend I'm researching."

"Ah, the Weaver of Whispers," the woman nodded. "A tale as old as time itself."

"Do you know more?" Amelia asked eagerly.

"Perhaps. But knowledge often reveals itself to those ready to receive it."

Amelia hesitated. "What do you mean?"

The woman gestured to the tapestry. "These threads contain more than mere fabric. They are a conduit, a link to something greater."

Leo chuckled quietly. "She's always this mysterious."

The woman cast him an amused glance. "Patience, Leo. Now, Amelia, would you care to share a cup of tea? We have much to discuss."

They sat at a small table by the window while the woman, who introduced herself as Seraphine, poured fragrant tea into delicate cups.

"The tapestry you found is part of a collection created by a craftswoman far ahead of her time," began Seraphine. "She wove pieces that connected different points of existence, threads capable of influencing the very fabric of reality."

Amelia listened intently, scepticism and fascination wrestling within her. "You're suggesting the tapestry has... capabilities?"

"Advanced methods may seem magical to those unfamiliar with them," Seraphine replied gently. "Tell me, have you experienced anything unusual since you found it?"

She recounted the strange events—the hum in the closed room, the shifting patterns, the faint glow. Seraphine nodded knowingly.

"The tapestry is responding to you," she said. "It senses your curiosity, your openness to possibilities."

Amelia felt a mix of excitement and apprehension. "What should I do?"

"Embrace it," Seraphine encouraged. "Let it guide you."

That night, back at the museum, Amelia stood before the tapestry. Taking a deep breath, she reached out and touched the hourglass symbol. The room seemed to ripple, and suddenly, she was no longer in the workshop.

She found herself in a vast expanse beneath a starlit sky. The air was crisp, filled with unfamiliar yet comforting scents. In the distance, a figure approached—a woman clothed in garments adorned with the same intricate patterns.

"Welcome, Amelia," said the woman, her voice a gentle echo. "I've been waiting for this moment."

"Who are you?" Amelia asked, her heart pounding.

"I am Elara," she replied. "Keeper of the threads."

Amelia felt a wave of calm wash over her. "Why am I here?"

"Because you have the potential to continue my work," said Elara. "The connexions that bind reality need attention."

"I don't understand," Amelia said, shaking her head.

"You will, in time," assured Elara. "Every action, every choice, weaves the tapestry of what is and what might be."

Before she could ask more, the scene gently faded, and Amelia was back in the workshop. The tapestry lay before her, its colours now more vibrant, the patterns subtly shifting as if alive.

Overwhelmed, she sat back, trying to make sense of what had happened. The rational part of her mind wrestled with the improbability of the experience, but deep inside, she felt a newfound calling.

In the days that followed, Amelia began to experiment, weaving small pieces inspired by the tapestry. She noticed subtle changes—a colleague finding inspiration after weeks of stagnation, a friend experiencing an unexpected opportunity. The events seemed minor but significant, suggesting a greater influence.

One afternoon, Leo visited her at the museum. "You seem different," he remarked.

"Do I?" she replied absently, her fingers deftly manipulating the threads.

"There's a spark about you," he said. "As if you've discovered a secret."

She smiled softly. "Perhaps I have."

He glanced at the tapestry. "You're still exploring its mysteries?"

"In a way," she said. "But it's more about creating new patterns now."

He nodded thoughtfully. "I wanted to thank you. The idea you gave me unlocked an important part of my research."

She looked up, surprised. "I didn't give you an idea."

He tilted his head. "You sent me that reference on interconnected systems."

The realisation dawned on her. The small weaving she'd left on her desk, adorned with intricate patterns—somehow, it had conveyed information.

"You're welcome," she said simply.

Over time, Amelia embraced her new role. She continued her work at the museum but spent her evenings weaving threads that subtly influenced the world around her. She stayed in contact with Seraphine, who guided her while encouraging her to explore on her own.

One night, as she was heading home, she noticed a man standing alone, looking lost and troubled. Instinctively, she reached into her bag and pulled out a small woven charm.

"Excuse me," she said softly. "I think this might be of help to you."

He looked at her, surprise in his eyes. "Why are you giving this to me?"

She smiled. "Sometimes, a small gesture can remind us we're not alone."

He accepted the charm hesitantly. As he held it, a subtle change came over him—the tension in his shoulders eased, a glimmer of clarity appeared in his eyes. "Thank you," he murmured.

She nodded and continued on her way, blending into the flow of passersby. Behind her, the man stood a little straighter, a new resolve visible in his posture.

Amelia understood that her journey was interwoven with those of others, each interaction a subtle adjustment in the complex tapestry of existence. No longer did she feel lost in the monotony of daily life; instead, she was part of something profound, collaborating silently with forces she was just beginning to comprehend.

One morning, she received a parcel at the museum—a book on advanced weaving techniques, filled with motifs and symbols both ancient and futuristic. Inside, a note: "For the Weaver of New Paths. Keep exploring.—S."

Amelia smiled, tucking the book under her arm. As she moved through the familiar halls, she felt a deep connexion to everything around her—the artefacts, the people, the very essence of the world. The tapestry that was once an enigma was now a gateway to infinite possibilities.

She stepped out into the bustling city, the cacophony of sounds now forming a harmonious symphony. Every face she passed bore untold stories, each moment a potential turning point. The world was filled with patterns waiting to be understood, threads waiting to be woven.

Amelia took a deep breath, feeling the pulse of life around her. With a renewed sense of purpose, she moved forward, ready to embrace the intricate design of what lay ahead.

Appendix 2: Detailed Hypothesis of Gödel's Theorem

The precise conditions of Gödel's theorems, though not overly demanding, are seldom presented in their most comprehensive and rigorous form.I have taken the trouble to sketch them here, in a form that is admittedly non-technical, yet simple and concise.The reader may find them in even more exhaustive detail in *The Internal Limitations of Formalisms*[21]:

1. The formal system is expressed in a first-order language.
2. It is consistent.
3. Its syntax can be expressed using recursive arithmetic.
4. It is at least as powerful as the PA system.

Condition 3 implies that it is possible to establish a one-to-one correspondence between the system's elements and the integers. Each constant, variable, function, and predicate must be assigned a distinct integer. Moreover, propositions must correspond to an integer that is a unique function of the integers associated with the elements they comprise. An example of such correspondence is Gödel encoding.

Additionally, the system must satisfy the following conditions: a) the property of a proposition being an axiom must be representable recursively; and b) the property of a proposition being the immediate consequence of one or more other propositions must be representable recursively.

Condition 4 can further be articulated through four clauses:

- The system must include a negation operator ¬ with its usual interpretation.

[21] Ladriere [110], pp. 160-163.

- The system must include the universal quantifier $\forall$ with its usual interpretation.
- The system must include an implication operator $\rightarrow$ with its usual interpretation. It must also have an inference rule that specifies the role of implication concerning the notion of derivability:
$$\frac{A \quad (A \rightarrow B)}{B}$$
- The system must incorporate a formalisation of recursive arithmetic, which means:

  a) it contains expressions unambiguously representing the integers;

  b) it includes symbols for variables, as well as function and predicate symbols with these variables as arguments;

  c) it incorporates a substitution operator that allows variables within a function or predicate to be replaced by expressions corresponding to integers;

  d) all recursive predicates in recursive arithmetic are representable in the formal system. In other words, any recursive predicate $P(x_1, ..., x_n)$ must correspond to a predicative expression $A(N_1, ..., N_n)$, where $N_1, ..., N_n$ are the system's expressions representing the integers $x_1, ..., x_n$.

Appendix 3: The Church-Turing Thesis

For further details on this topic, the reader is referred to Chapter 6 of *The Internal Limitations of Formalisms*[22], which inspired this appendix.

Effective Method

Shortly after the emergence of the incompleteness theorems, the question arose whether there existed an "effective method" to determine the provability of any given proposition. This inquiry defines the *resolvability* of a formal system.

Définition 29. *A theory is termed resolvable (in the syntactic sense) if there exists an "effective method" allowing one to decide, for any proposition ϕ of the language, whether ϕ is derivable or not. The problem of finding such a method is known as the decision problem*[23].

What is meant by "effective function" can be interpreted in various ways. No fewer than three notions have been proposed to capture this intuition[24]. The Church-Turing Thesis posits that these approaches are equivalent and indeed capture the concept of "computability". In other words, everything that is computable can be expressed through one of these three methods, and vice versa.

1. Turing-computable function
2. General recursive function
3. λ-definable function

[22] Ladriere [110], pp. 227-278.
[23] Ladrière, *Ibidem*, p. 61.
[24] Ladrière, *Ibidem*, p. 231-235.

Turing-computable function A Turing machine includes a tape, a reading head, an execution unit, and memory. The tape, assumed to be infinitely long, is divided into cells. Each cell contains a symbol from a finite alphabet or a "0" denoting an empty cell. The reading head deciphers the symbol on the current cell. It can move left or right along the tape, based on instructions from the execution unit. It may also erase the symbol in the cell or print a new one. The memory holds the current state of the Turing machine and the programme. This programme is presented as a table that, for each possible state of the machine, defines the instruction sequence to execute.

The "input" of a Turing machine is the initial state of the tape before the programme starts, and "output" is the state of the tape when the machine halts. A function f is Turing-computable if it can be described by a Turing machine. Any element x of its domain can then be represented by a sequence of characters and input into the machine. After a finite set of operations, the machine will halt on an output representing $f(x)$.

The universal machine is a peculiar Turing machine capable of simulating the behaviour of any other Turing machine. It takes two character strings, m and x, as input, corresponding to the description of a Turing machine M and an admissible input x for that machine. The universal machine then outputs $M(x)$.

General recursive function Initially, the notion of "computable function" was approximated by the idea of *primitive recursive function*. This was defined recursively from functions of one variable using various manipulations, such as function composition and the successor operator.

It was later realised that some recursive functions, like the Ackermann function, did not satisfy this definition. Thus, the definition of a *general recursive function* was introduced. A function $f : D \subseteq \mathbb{N} \to \mathbb{N}$ is *general recursive* if the image of every $n \in D$ can be computed by an algorithm. In other words, if a finite set of instructions exists such that, after a finite number of well-defined steps, the image $f(n)$ is obtained. When $D = \mathbb{N}$, it is termed a *partial recursive function*. Kleene formalised this notion.

λ-definable function In a first-order formal system, symbols exist to designate variables, constants, functions, and predicates. Functions do not follow the same syntactic rules as variables: for instance, they cannot be introduced as arguments in other functions or predicates. This creates a complex hierarchy among variables. Additionally, this type of language allows one to refer to $f(x)$ without defining the calculation procedure for finding the image.

Church's idea was to construct a system involving only one type of variable, in addition to a special symbol denoted λ. This symbol allows functions to be constructed by explicitly defining the operations to perform on

the variables.

Formally, for any expression y in this system, the formula $\lambda x.y$ denotes a function that maps every x to y. In this formula, x is a bound variable, while y can be either a free variable or another formula.

By way of illustration, the identity function is denoted $\lambda x.x$, while natural numbers are constructed as follows (interpreting y as representing "0" and x as the "successor" operator):

$$\begin{aligned} N_1 &\equiv \lambda xy.xy \\ N_2 &\equiv \lambda xy.x(xy) \\ N_3 &\equiv \lambda xy.x(x(xy)) \\ &\dots \end{aligned}$$

In this context, a function f is said to be λ-definable if it can be represented in λ-calculus. More formally, if N_a and N_b represent integers a and b, there must exist a formula F in lambda calculus such that:

$$f(a) \xrightarrow[\text{calculation}]{} b$$

$$FN_a \xrightarrow[\text{conversion}]{} N_b$$

Where "conversion" denotes a sequence of transformations allowed by the syntactic rules of the λ-calculus (substitution, reduction, and expansion).

Turing's Theorem

Kleene demonstrated the equivalence between general recursive functions and certain types of λ-definable functions. Turing, in turn, showed the equivalence between these two notions and his own definition of a computable function.

Theorem 9. *There is equivalence between*[25]*:*

1. *General recursive function.*
2. *Turing-computable function.*
3. *λ-definable functions defined over the entire $\mathbb{N}^n$, where n is the number of variables.*

It is possible to include λ-definable functions whose domain is $D^n \subset \mathbb{N}^n$. In this case, the equivalence becomes:

[25] Ladrière, *Ibidem*, p. 236.

Theorem 10. *There is equivalence between*[26]*:*

1. *Partial recursive function.*
2. *λ-definable function.*

Each of these three definitions of computability can thus be used to attempt to solve the decision problem. Church addressed this problem with the notion of λ-definability, Kleene with that of general recursiveness, and Turing with his own. They all achieved very similar results.

Theorem 11. *No theory satisfying the hypotheses of Gödel-Rosser's incompleteness theorems is resolvable*[27]*.*

Thus, there is no universal algorithm for determining the provability of a proposition in such theories. The proof method used in this theorem, for each of the three definitions of computability, employs a diagonal argument.

[26] Ladrière, *Ibidem.*
[27] Ladrière, *Ibidem*, p. 238-244.

Appendix 4: Cantor's Paradox

Cantor's paradox, proven in 1899, declares that there is no set of all sets; or more precisely, no cardinal of all cardinals exists. The notion of cardinal was introduced by Cantor to rank sets according to their sizes, particularly infinite ones. Contrary to what one might think, infinity is not a singular, unchanging entity. There are "larger" infinities and "smaller" infinities. For example, the set $\mathbb{N}$ of natural numbers is smaller than the set $\mathbb{R}$ of real numbers, even though both contain an infinite set of numbers.

To understand how one infinite set can be "larger" than another, I will first present three definitions of a set's "size," including the concept of cardinality used by Cantor. I will then demonstrate how these ideas can show that the set of all sets is never large enough to contain the set of its subsets. This will require an introduction to an argument known as "the diagonal argument."

How to Define the Size of an Infinite Set?

It seems natural to think that the "size" of a set is given by the number of elements it contains. The set $E_1 = \{2, 4, 6, 8\}$ is larger than the set $E_2 = \{2, 4, 8\}$ because it contains one more element. According to this definition, it follows naturally that the set E_1 will always have a size greater than or equal to any of its subsets. A subset of E_1 is obtained by selecting one or more elements from E_1. For example, $\{2, 8\}$ and $\{4, 6, 8\}$, as well as the empty set $\{\}$ and the set E_1 itself.

In the presence of infinite sets, counting the number of elements becomes difficult, making the notion of size harder to define. To illustrate, consider the following two sets:

$$A = \{2, 4, 6, 8, 10...\}$$

$$B = \{2, 4, 8, 16, 32...\}$$

The question is, which of these two sets is "larger," and in what exact sense? At first glance, both have an infinity of elements, but does that mean they are the same size? I would like to discuss three distinct approaches: one using the notion of inclusion, another based on density, and a third resting on the concept of cardinality.

Approach 1: Inclusion

Initially, we observe that the set B of powers of two is included in the set A of even numbers, denoted $B \subseteq A$. We might think B is "smaller" than set A. To formalise this intuitive notion of "size," we could propose the following definition:

Définition 30. *Let A and B be two sets. We define:*

- $A \prec B$ *if and only if* $A \subset B$.
- $A \preceq B$ *if and only if* $A \subseteq B$.
- $A \approx B$ *if and only if* $B \subseteq A \subseteq B$.

The symbols $\prec, \preceq$ and $\approx$ in this definition are read respectively as "is smaller than," "is smaller or equal to," and "is equal to." Identifying the symbol "$\preceq$" with "$\subseteq$" is not merely metaphorical. It can be shown that the essential algebraic properties expected of the relation "is algebraically smaller or equal to" ("$\preceq$") are satisfied by the relation "is a subset of" ("$\subseteq$"). We have, for instance:

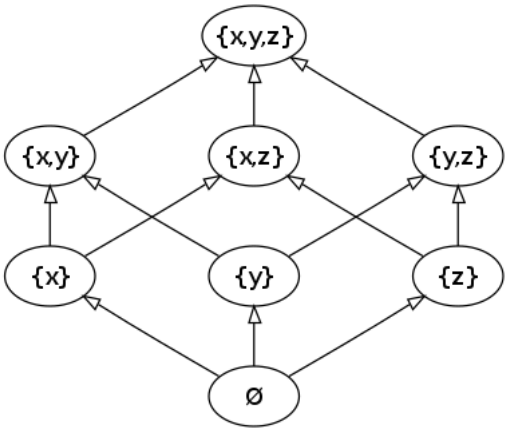
$$\begin{aligned} &A \preceq A && \text{(reflexivity)} \\ &\text{if } A \preceq B \text{ and } B \preceq A, \text{ then } A \approx B && \text{(antisymmetry)} \\ &\text{if } A \preceq B \text{ and } B \preceq C, \text{ then } A \preceq C && \text{(transitivity)} \end{aligned}$$

And similarly for the subset relation:

$$\begin{aligned} &A \subseteq A && \text{(reflexivity)} \\ &\text{if } A \subseteq B \text{ and } B \subseteq A, \text{ then } A = B && \text{(antisymmetry)} \\ &\text{if } A \subseteq B \text{ and } B \subseteq C, \text{ then } A \subseteq C && \text{(transitivity)} \end{aligned}$$

These three properties define what is known in mathematical logic as a *partial order* on the class of sets. They allow us to compare and rank sets against one another. Below is an illustration using sets containing three elements x, y, and z.

The Hasse diagram ?? provides a ranking of the subsets of $\{x, y, z\}$ using the order relation "$\subseteq$." The empty set "$\emptyset$" is at the bottom because it is a



subset of all other sets, while $\{x, y, z\}$, which is contained in no other set of $\mathbb{P}(\{x, y, z\}$, appears at the top. We observe that the sets $\{x\}, \{y\}$, and $\{z\}$, each with only one element, are situated at the first level. Similarly, the two-element sets $\{x, y\}, \{x, z\}$, and $\{y, z\}$ are at the second level. There are no arrows between sets at the same level because these sets are incomparable to each other. Indeed, we have $\{x\} \not\subseteq \{y\}$ and $\{y\} \not\subseteq \{x\}$.

This example illustrates that not all sets are comparable. This can occur even when sets do not have the same number of elements. For example, the set $E = \{1, 2, 3\}$ and $E' = \{1, 3, 4, 5\}$ are not comparable because $E \not\subseteq E'$ and $E' \not\subseteq E$. Yet, we would like to be able to say that the set E, as it contains more elements, is larger than E'. Thus, a different definition is needed.

Approach 2: Density

Rather than asking whether one set is included in another, which is relatively rare, one could reason as follows: both sets A and B indeed contain an infinity of integers, but they also lack an infinity of integers. The set A of even numbers contains exactly half of the natural numbers, as it lacks odd numbers. The set B of powers of two contains only a fraction of even numbers; it is missing odd numbers, as well as most even numbers that are not powers of two, like $6, 10, 12, 14, 18, 20$, etc. We would therefore say that set A is more *dense* than set B.

To formalise this notion of *density*, we can investigate the probability that, if an integer is randomly chosen, it belongs to A or B? Suppose, to begin with a simple case, that the integer chosen at random is between 0 and 100. It follows a homogeneous probability distribution, meaning that there is one out of one hundred chance of choosing each number, 0, 1, etc.

$$A = \{\underbrace{2, 4, 6, ..., 98,}_{\text{49 even numbers}} 100, 102, ...\}$$
$$B = \{\underbrace{2, 4, 8, 16, 32, 64,}_{\text{6 powers of two}} 128, 256, ...\}$$

In the interval $[0, 99]$, there are exactly 49 even numbers other than zero and 6 powers of two greater than one. Thus there is approximately a 50% chance that a number belongs to A and a 5% chance that it is in B. This supports the idea that A is denser than B. However, it is arbitrary to pick an integer from the interval $[0, 99]$ rather than $[0, 999]$ or $[0, 9999]$. Hence, we are interested in the behaviour of these probabilities as the interval becomes $[0, n]$ when the boundary n tends to infinity.

If we increase the size of the interval $[0, 99]$, the number of even natural numbers grows linearly, while the number of powers of two follows a decreasing logarithmic curve. Note that, in the tables below, the notation $\lfloor . \rfloor$ denotes rounding down to the nearest integer (e.g., $\lfloor \log_2(3) \rfloor = \lfloor 1.58496 \rfloor = 1$).

Interval	$[1, 100]$	$[1, 200]$	$[1, 300]$	...	$[1, n]$
# Of integers in B	50	100	150	...	$\lfloor n/2 \rfloor$
Probability	50%	50%	50%	...	50%

Interval	$[1, 100]$	$[1, 200]$	$[1, 300]$	...	$[1, n]$
# Of integers in A	6	7	8	...	$\lfloor \log_2(n) \rfloor$
Probability	6%	3.5%	2.6%	...	?

To arrive at a precise result, we seek formulas P_A and P_B giving the probability that a randomly chosen integer from the interval $[0, n]$ lies in set A and B, respectively. This probability is given by:

$$\begin{aligned} P_A(n) &= \frac{\text{\# of even naturals in the interval}}{\text{size of the interval}} \\ &= \frac{\lfloor \frac{n}{2} \rfloor}{n} \end{aligned}$$

and

$$\begin{aligned} P_B(n) &= \frac{\text{\# of powers of two in the interval}}{\text{size of the interval}} \\ &= \frac{\lfloor \log_2(n) \rfloor}{n} \end{aligned}$$

It is easy to see that as the size of the interval $[0, n]$ tends towards infinity, the probability $P_B(n)$ tends to zero, while the probability $P_A(n)$ tends to $1/2$. Thus, under such consideration, for most naturals, they are not typically powers of two, which reinforces the observation that set A is infinitely denser than set B.

We can formalise the notion of "size" between arbitrary sets A and B like so:

Définition 31. *Let A and B be two sets. We define:*

- $A \prec_\delta B$ *if and only if* $\lim_{n\to\infty} P_A(n) < \lim_{n\to\infty} P_B(n)$.
- $A \preceq_\delta B$ *if and only if* $\lim_{n\to\infty} P_A(n) \leq \lim_{n\to\infty} P_B(n)$.
- $A \approx_\delta B$ *if and only if* $\lim_{n\to\infty} P_A(n) = \lim_{n\to\infty} P_B(n)$.

I've used the notation $\prec_\delta$, with a small δ in the subscript to differentiate this approach from the one used before, based on inclusion. It is possible to show that $\prec_\delta$ forms an *order relation* and thus respects the algebraic properties of *reflexivity*, *antisymmetry*, and *transitivity* as previously observed.

The advantage of this new definition of "size" is that it allows comparison between sets that were incomparable when inclusion was used ("$\subseteq$"). For instance, considering the set of parts of $\{x, y, z\}$, can $\{x, y\}$ and $\{x, z\}$ be compared in terms of density? Yes, because the probability of a randomly chosen element e in $\{x, y, z\}$ being in $\{x, y\}$ is equal to the probability of it being in $\{x, z\}$. In terms of density, the two sets thus have the same size, although they aren't included in each other. Yet, we will see that this is not the only admissible measure of a set's "size."

One issue with this density approach arises in that, to be comparable, sets A and B must both be subsets of a common set. In our example, this common set is $\mathbb{N}$. One could also choose $\mathbb{R}$ or $\mathbb{C}$, which would yield different densities (the probability that a real number $r \in \mathbb{R}$ randomly chosen is a natural even number is zero).

One might wonder, however, what happens when sets A and B do not belong to a common set? In such a case, they are simply not comparable using densities. To illustrate, consider the set $A' = \{a_1, a_2, ..., a_n\}$ and $B' = \{b_1, b_2, ..., b_m\}$, where $a_1, ..., a_n$ are letters from the alphabet $\mathcal{X}_a$ (e.g., the Roman alphabet), and $b_1, ..., b_m$ are letters from a distinct alphabet $\mathcal{X}_b$ (e.g., the Russian alphabet). It is, however, possible to bypass this difficulty by noting that A' and B' are both inside $\mathcal{X}_a \cup \mathcal{X}_b$.

Approach 3: Cardinality

I will now present an argument that contradicts results obtained using the previous two methods, which suggested that the set A of even numbers and

the set B of powers of two would have the same size. My argument concerns not inclusion or density but what is known as cardinality.

Cardinality can be understood by reasoning as follows. Suppose we have two large sacks filled with objects. To determine which sack is larger, we randomly remove an object from the first sack, then from the second. The sack emptied first is the smallest. If they empty simultaneously, they are the same size. In the case of our infinite sets A and B, we might intuitively believe they are the same size since neither will ever empty. How can we rigorously show this property? How can we determine cases where one infinite set "empties" faster than another?

Notice that it is possible to map set A to the set of natural numbers $\mathbb{N}$:

$$\begin{array}{lccccccc} \mathbb{N}: & 0 & 1 & 2 & 3 & ... & & n \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & & \downarrow \\ \text{A}: & 2 & 4 & 6 & 8 & ... & & 2(n+1) \end{array}$$

This table implies that each natural number $n \in \mathbb{N}$ corresponds to a unique element $a \in A$, and conversely, each element $a \in A$ corresponds to a unique natural number $n \in \mathbb{N}$. This type of relationship is called a *bijection* of A into $\mathbb{N}$.

Intuitively, the existence of a bijection indicates there are "as many" naturals in $\mathbb{N}$ as there are elements in A: each time an element is removed from the set of naturals, one can be removed from set A; and vice versa. Thus, from this viewpoint, the two sets are the same "size." In technical terms, A and $\mathbb{N}$ are **countably equivalent**, or they have the same **cardinality**.

Now, the question is whether B and $\mathbb{N}$ are also countably equivalent? The answer is yes. To see this, consider the following correspondence:

$$\begin{array}{lccccccc} \mathbb{N}: & 0 & 1 & 2 & 3 & ... & & n \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & & \downarrow \\ \text{B}: & 2 & 4 & 8 & 16 & ... & & 2^{n+1} \end{array}$$

Since A is equivalent to $\mathbb{N}$ and $\mathbb{N}$ is equivalent to B, it can be demonstrated that A is also equivalent to B. One merely has to consider the bijection below:

$$\begin{array}{lccccccc} \mathbb{N}: & 2 & 4 & 6 & 8 & ... & & 2n \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & & \downarrow \\ \text{B}: & 2 & 4 & 8 & 16 & ... & & 2^n \end{array}$$

These findings suggest that A and B have the same "size" (in terms of cardinality), even though (1) there are infinitely more even numbers than there are powers of two (in terms of density), and (2) the set B is a subset of A. In Cantor's notation, A, B, and $\mathbb{N}$ are said to be equipotent, and their size is denoted by the symbol $\aleph_0$.

Given these observations, we are led to wonder whether sets exist that are not equipotent to $\mathbb{N}$? Cantor's theorem, which we will examine later, affirms this and shows that:

- The set $\mathbb{R}$ has a cardinality strictly greater than the cardinality of $\mathbb{N}$, noted $|\mathbb{R}| > \aleph_0$;
- The set of parts of $\mathbb{R}$ has a cardinality strictly greater than $\mathbb{R}$'s, noted $|\mathcal{P}(\mathbb{R})| > |\mathbb{R}|$;
- The set of parts of parts of $\mathbb{R}$ has a cardinality strictly greater than the parts of $\mathbb{R}$, noted $|\mathcal{P}(\mathcal{P}(\mathbb{R}))| > |\mathcal{P}(\mathbb{R})|$;
- etc.

At this point, a meticulous reader might ask what it means for a set to have a cardinality strictly greater than another's? Until now, we have defined what it means for two sets to have the same cardinality, but an order relation has yet to be defined.

To answer this question, consider the sets $A' = \{x, y\}$ and $B' = \{x, y, z\}$, noticing that a bijection between A' and B' is impossible, so they do not share the same cardinality. However, a *total injection* can be made from A' into B' and a *surjection* from B' into A'. Due to these relations, it is said that the cardinality of B' is greater than that of A.

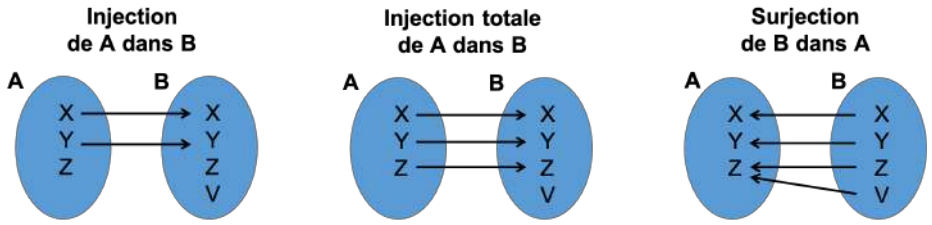


Figure 3: The figure on the left illustrates an injection of A into B. This transformation ensures no two elements of A have the same image. The figure in the middle shows a total injection, meaning the transformation is injective, and additionally, all elements of A have an image in B (denoted $A \preceq_\alpha B$). Finally, the graph on the right depicts a surjection from B' into A'. This transformation ensures every element of A' is an image of an element of B' (denoted $B \succeq_\alpha A$). Note that, for the figure on the left, A and B cannot be compared because some elements of A lack an image in B, and certain elements of B are not images of elements in A (neither $A \preceq_\alpha B$, $B \succeq_\alpha A$, nor $A \approx_\alpha B$ can be written).

The following definition is suggested:

Définition 32. *Let A and B be two sets. We define:*

- $A \approx_\alpha B$ *if and only if a bijection of* A *into* B *exists.*
- $A \preceq_\alpha B$ *if and only if a total injection of* A *into* B *exists.*
- $A \succeq_\alpha B$ *if and only if a surjection of* A *into* B *exists.*

The symbols $\approx_\alpha$, $\preceq_\alpha$, and $\succeq_\alpha$, where a small "α" is placed in the subscript, once again form an *order relation* and uphold the properties of *reflexivity*, *antisymmetry*, and *transitivity*[28].

To say of the relation $\preceq_\alpha$ that it forms an order relation implies, in particular, that if $A \preceq_\alpha B$ and $B \preceq_\alpha A$, then $A \approx_\alpha B$. In plain English: if a total injection exists from A into B, and a total injection exists from B into A, then A and B share the same cardinality. This significant theorem was given its proof by Bernstein and Cantor around 1896[29].

Counterintuitive Properties of Infinity

After exploring these three approaches — inclusion, density, and bijection — we find that several properties once intuitive in a finite world must be rejected to remain coherent in the presence of infinite sets:

- **Size can be defined in multiple ways:** There are various ways to define a set's "size." One relies on inclusion, another on density, and another on the concept of equipotence. In finite contexts, these definitions coincide and can be interchanged. However, this ceases to be the case when *actual infinity* is accepted.
- **The whole is not larger than its parts:** A mereological principle, which Aristotle fervently defended, asserts that the whole is greater than the parts. Bolzano first proposed abandoning this principle to resolve difficulties arising from the existence of *actual infinite* sets. Later mathematical developments confirmed this notion. Specifically, within Cantor's theory, a bijection can be established between an infinite set and an infinite part of that set, as we've seen with sets A and B. This demonstrates that *cardinality* is distinct from *inclusion.*
- **A zero-probability event is not impossible:** To see this, consider: what is the probability that, if a natural number is chosen at random

[28] Antisymmetry expresses that if a total injection exists between A and B (denoted $A \preceq_\alpha B$) and if a total injection also exists between B and A (denoted $B \preceq_\alpha A$), then a bijection exists between A and B (denoted $A \approx_\alpha B$). This theorem, proven by Cantor and Bernstein, is far from trivial, but it is essential for making countable equinumerability a valid measure of a set's size.

[29] Georg Cantor stated this theorem without proof in 1887, then proved it in his work *On the Foundations of Transfinite Set Theory* (1895-1897), using ordinal theory. Felix Bernstein, one of Cantor's students, outlined another proof at just 18 years of age in 1896. Published on Cantor's recommendation in *Lectures on the Theory of Functions,* it appeared under Émile Borel's pen in 1898.

(equiprobably), it is a power of two? The calculation presented in exercise (c) shows that the probability that an integer m randomly chosen from $\{1, 2, ..., n\}$ is a power of two tends toward zero as n approaches infinity. We conclude that the probability of an integer m drawn from $\mathbb{N}$ being a power of two is zero. Although challenging to conceptualise, how can m not be $2, 4, 8, 16$, etc., when these numbers are valid candidates?

To resolve this apparent paradox, one might argue that the probability is not *zero*, but *infinitesimal*. This implies that $\mathbb{N}$ constitutes an *potential infinity*. However, Cantor's idea was to assign $\mathbb{N}$, considered as a whole, a determined size, denoted $\aleph_0$. Instead of viewing $\mathbb{N}$ as a constructively built (and recursively formed) set consisting of $\{1\} \cup \{2\} \cup ... \cup \{n\} \cup ...$, it is seen as a whole given at once. Rather than being seen as the product of an infinite enumeration, such as $\lim_{n\to\infty}\{1, 2, ..., n\}$, it is viewed as a genuine *actual infinity*.

Since $\mathbb{N}$ forms an actual infinity, the probability of a randomly (equiprobably) drawn natural number m being a power of two is not infinitesimal; it is strictly zero. Thus, mathematical possibility must be distinguished from probabilistic possibility. An event that is mathematically contradictory (like drawing a natural number both even and odd) always has a zero-probability. However, a non-contradictory event (like drawing a natural number that is a power of two) does not always have a non-zero probability. Accepting the existence of an actual infinity, effectively acknowledging a distinction between the realm of possibilities and the realm of probabilities.

If the distinction between "logical impossibility" and "probabilistic impossibility" is still unclear, I invite you to participate in a little game. I'm going to think of a real number between 0 and 1. Then I'll ask you to give me the number you choose. If it matches the number in my mind, you will receive 20/20 on the exam. Otherwise, nothing will happen. The question is: do you accept to play the game? One might be tempted to agree since, at its core, nothing is lost by participating. However, your chances of winning are strictly zero. Why is that? Because a real number is constructed with an infinite number of decimals. Even if you manage to guess ten decimals of the number I have in mind, what are the odds that you also guess the next ten? The following hundred? The next thousand? The $\aleph_0$?

Cantor's Diagonal Argument

Now that we have clarified some peculiarities of infinite sets, I would like to introduce Cantor's diagonal argument. This argument is a key ingredient in

proving that there is no set of all sets.

The Set of Subsets

Cantor's diagonal argument demonstrates that any set E has a size strictly smaller than that of its set of subsets, denoted $\mathbb{P}(E)$. To recall, the set of subsets is constructed by enumerating all subsets possible using a given original set. For example, the set $\{1, 2\}$ admits four subsets: $\{\}, \{1\}, \{2\}, \{1, 2\}$. It is straightforward to show that if the set E contains n integers, then the set $\mathbb{P}(E)$ will have 2^n.

The question raised is what happens to the size of $\mathbb{P}(E)$ when E is an infinite set? The diagonal argument suggests that the set $\mathbb{P}(E)$ is always larger in cardinality than the original set E. Thus, if we use the notion of size formalised by bijection, we have $E \preceq_\alpha \mathbb{P}(E)$.

To address this question, let's suppose E and $\mathbb{P}(E)$ are of the same size. This means we could create a book, with a countable infinity of pages, each containing a distinct set of naturals. Page one could contain the set $\{1, \{2, 3\}\}$, page two could have the set $\{\{7\}\}$, and so on. All pages are distinct, yet the entire book enumerates all possible sets of naturals.

This book exemplifies a bijection between the set of integers and the set of subsets. Each page of the book corresponds to an integer: page one carries the number 1, page two bears the number 2, etc. In addition, each page lists a subset of integers. There is thus a bijection between $\mathbb{N}$ and $\mathbb{P}(N)$.

However, I'd like to show that this book is paradoxical and cannot exist. To do so, let's define an "ordinary number" as an integer n not represented on the n-th page. The number 2 is ordinary for instance, since the set on page number 2, namely $\{\{7\}\}$, does not include the number 2. However, the number 1 is not because page number 1 contains 1 in the set it represents: $\{1, \{2, 3\}\}$.

If Cantor's book contains all possible sets of naturals, there must be a page, numbered n', containing the set of all ordinary numbers and no other. Therefore, the set on page n' would include the number 2 but exclude the number 1. Do you follow? The question to ask is: is n' an ordinary number?

Suppose n' is an ordinary number. It must then appear on the page listing all ordinary numbers. However, that page is numbered n'. This contradicts the definition of an ordinary number. If, instead, n' is not ordinary, by definition, it must appear on the page numbered n'. Yet this page should only list ordinary numbers, contradicting the status of n' as ordinary. Thus, the book of Cantor is paradoxical.

To avoid contradictory mathematics, the existence of such a book must be disallowed. Logically, this implies that it is impossible to enumerate all sets of natural numbers by listing them in an infinite book. Formally, one can say that no *bijection* exists from $\mathbb{N}$ into $\mathcal{P}(\mathbb{N})$. However, it is possible

to create a *total injection* from $\mathbb{N}$ into $\mathcal{P}(\mathbb{N})$. To achieve this, one must associate each natural number n with a subset of naturals such that two different naturals n_1 and n_2 do not share the same image. An example of total injection is given below:

$$\begin{array}{rccccccl} \mathbb{N}: & 1 & 2 & 3 & ... & n & & \text{(page numbers)} \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & \downarrow & \\ \mathcal{P}(\mathbb{N}): & \{1\} & \{2\} & \{3\} & ... & \{n\} & & \text{(distinct sets of naturals)} \end{array}$$

This example corresponds to an infinite book where each page numbered n contains the unique singleton $\{n\}$. Notice this book would not contain an ordinary number, and no page would bear a set of ordinary numbers. Thus, there would be no paradox, meaning our total injection from $\mathbb{N}$ into $\mathcal{P}(\mathbb{N})$ is valid, and $\mathbb{N} \preceq_\alpha \mathcal{P}(\mathbb{N})$. Since no bijection of $\mathbb{N}$ into $\mathcal{P}(\mathbb{N})$ exists, we conclude the inequality is strict: $\mathbb{N} \prec_\alpha \mathcal{P}(\mathbb{N})$.

Generalisation to Any Set

The logic presented here can be generalised to any set E, whether finite or infinite, whether containing numbers, letters, or other objects. It remains true even for infinite sets of greater cardinality than $\mathbb{N}$, such as $\mathbb{R}$ or $\mathcal{P}(\mathbb{R})$. I will sketch the proof showing $E \prec_\alpha \mathcal{P}(E)$ for any set E. The method proceeds in two stages. First, we demonstrate that a *total injection* exists from E into $\mathcal{P}(E)$:

$$\begin{array}{rcccccc} E: & x_1 & x_2 & x_3 & x_4 & ... & x_n \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & \downarrow \\ \mathcal{P}(\mathbb{N}): & \{x_1\} & \{x_2\} & \{x_3\} & \{x_4\} & ... & \{x_n\} \end{array} \quad \begin{array}{l} \text{(elements of E)} \\ \\ \text{(sets of E)} \end{array}$$

This allows us to say $E \preceq_\alpha \mathcal{P}(E)$. To show the inequality is strict, it remains to demonstrate no *bijection* from E into $\mathcal{P}(E)$ exists. To achieve this, we assume such a bijection exists and show it leads to a contradiction. Let this bijection be f:

$$\begin{array}{rcccccc} E: & x_1 & x_2 & x_3 & x_4 & ... & x_n \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & \downarrow \\ \mathcal{P}(\mathbb{E}): & f(x_1) & f(x_2) & f(x_3) & f(x_4) & ... & f(x_n) \end{array} \quad \begin{array}{l} \text{(elements of E)} \\ \\ \text{(images)} \end{array}$$

We consider the set $C = \{x \in E | x \notin f(x)\}$. In plain English, the set C is the set of elements $x \in E$ such that x is not contained in the set mapped by $f(x)$. In the special case where $E = \mathbb{N}$, the set C designates nothing other than the set of ordinary numbers.

Notice that C is a subset of E. Thus, there exists a $c \in E$ such that $f(c) = C$ (since we assume f is a bijection from E into $\mathcal{P}(E)$). We then pose

the question: is c within C? If $c \in C$, it implies $c \notin f(c)$ (by definition of C), leading to the contradiction: $c \notin C$ (since $f(c) = C$). Conversely, if $c \notin C$, it implies $c \in f(c)$ (by definition of C), hence the contradiction: $c \in C$ (since $f(c) = C$). We thus conclude C is paradoxical, and the bijection f cannot exist. Consequently, we have $E \prec_\alpha \mathcal{P}(E)$.

To summarise in simpler terms, imagine you have a box E' containing all the pralines in the world: blue, red, yellow pralines, etc. You wonder how many different boxes of pralines can be conceived by picking pralines from this initial box. You might create a box containing only the navy blue and opalescent red pralines, excluding all others; or a box containing only dark pralines; only light pralines; or solely blue pralines, etc. You can also create praline boxes that contain other praline boxes, which in turn contain other praline boxes, and so forth. Cantor's diagonal argument shows that the box containing all possible praline boxes imaginable will be so large that it will never fit in the box E' containing one example of every praline in the world. This is what Cantor's theorem teaches us.

Cantor's Paradox

For any set E, I have shown that $\mathbb{P}(E)$ has a size strictly larger than E. Applying this theorem, we now show that the set of all sets, denoted ENS, is paradoxical. Firstly, the fact that ENS is the set of all sets means it includes the set $\mathbb{P}(ENS)$. So we have $\mathbb{P}(ENS) \in ENS$. Moreover, ENS also includes the set of its subsets, denoted $\mathbb{P}(ENS)$, and thus we can trivially construct a total injection of $\mathcal{P}(ENS)$ into ENS:

$$\begin{array}{rccccccl} \mathcal{P}(ENS): & x_1 & x_2 & x_3 & x_4 & ... & x_n & \text{(elements of } \mathcal{P}(ENS)\text{)} \\ & \downarrow & \downarrow & \downarrow & \downarrow & ... & \downarrow & \\ ENS: & x_1 & x_2 & x_3 & x_4 & ... & x_n & \text{(images)} \end{array}$$

This leads us to conclude that $\mathcal{P}(ENS) \preceq_\alpha ENS$. However, given Cantor's theorem, we can also write $ENS \prec_\alpha \mathcal{P}(ENS)$, which leads to a paradox. If mathematical non-contradiction is to be preserved, the existence of the set ENS must be rejected. This result is disconcerting, as the reason for being forbidden to speak of the set containing all possible sets is not *a priori* obvious.

At the time of Cantor's paradox publication (1899), mathematicians did not seriously regard the threat this theorem posed to the axiomatisation of mathematics. Cantor's theory was viewed as dubious, even dangerous, for allowing discussion of actual infinite sets. Instead of attempting to reconstruct set theory, it was easier to collectively dismiss ordinal and cardinal theories. Prominent figures like Hilbert insisted that mathematics adhere to a rigorous approach, constructing each mathematical object explicitly through a finite number of operations. This requirement *de facto* rules out

considering $\mathbb{N}$ as an actual infinity. All that can be discussed is the set $\{1, 2, 3, ..., n\}$, where n is a finite natural number, as large as desired.

Rather than focusing on abstract and potentially contradictory objects, Hilbert considered the *sign* the only admissable definition of a mathematical object. Far from an ideal platonic entity or pure intuition, the sign is most tangible, existing on paper, following strict syntactical rules, and subject to traceable and verifiable manipulations. Of course, Hilbert was aware that, in doing mathematics, it is sometimes necessary to use abbreviations or not completely formalised reasoning.

Hence, he distinguished metamathermatics, whose role was to demonstrate the consistency of usual mathematical operations, from mathematics proper. Within metamathermatics, everything needed to be derivable from axioms using a finite number of signs. At this point, Hilbert imagined the fiction of a mathematician equipped with a pencil and a stack of finite-sized papers, expandable to any desired extent. Could this mathematician derive all known mathematical results using a finite number of signs? Could all 23 unresolved mathematical problems at the dawn of the twentieth century be resolved?

Hilbert presented this challenge to the scientific community during the 1900 International Congress of Mathematicians, held in Paris, gathering experts from all disciplines. Curiously, the envisaged calculator from Hilbert imagined prefigures the computer. Indeed, in English, the term "computer" long referred to a person capable of deriving paper-based calculations manually. Only after the development of early computers did the term gradually shift to its current meaning.

In this vibrant atmosphere, where Hilbert's formalism and the axiomatic pursuit permeated the epistemic landscape, it took Russell's paradox in 1902 for the fragility of naïve set theory to be recognised universally. This new paradox's strength was that it required no assumptions about the existence of ordinals or cardinals. It challenged both Frege's *Foundations of Arithmetic* and Hilbert's dream, leading to new axiomatic approaches: Russell's type theory, Zermelo-Fraenkel theory, and Lesniewski's "classless" ontology.

Appendix 5: Transfinite Numbers

The definition of ordinal numbers used here is von Neumann's, which is more easily formalised in set theory. An ordinal α is a transitive set for which the membership relation $\in$ defines a strict order relation. If $x \in \alpha$, then $x \subset \alpha$ and if $x, y \in \alpha$, then either $x \in y$, or $x = y$, or $y \in x$. Furthermore, all the elements of an ordinal are themselves ordinals. We shall sometimes write $x < y$ instead of $x \in y$ and $x \leq y$ for $x \subset y$. The successor operation is denoted s, and for any ordinal α, it satisfies $s(\alpha) = \alpha \cup \alpha$. In particular, finite ordinals are identified with natural numbers. Zero is the empty set and, by applying the operation s recursively, we define $n + 1 = n \cup n = 0, 1, ..., n$. The set of all finite ordinals is ω_0, also denoted ω. It is the smallest infinite ordinal, coinciding with $\mathbb{N}$. The first uncountable ordinal is ω_1. The second uncountable ordinal not writable as an operation on lower ordinals is ω_2. Continuing in this manner, we define $\omega_3, \omega_4, ..., \omega_{\omega_0}$, and so on.

For any ordinal α, we denote $\cup(\alpha)$ the union of all elements of α. Either α has a maximal element $\beta = \cup(\alpha)$ and then we have $s(\beta) = \alpha$, or it does not, and $\cup(\alpha) = \alpha$. In the first case, α is said to be a non-limit ordinal. In the second case, it is a limit ordinal and we denote $\alpha \in \text{Lim}$. Ordinals of the form ω_α are all limit ordinals.

The addition "+" and multiplication "$\cdot$" for ordinals are defined by a set of axioms. Given two ordinals α and β:

$$\begin{array}{ll} \alpha + 0 & = \alpha \\ \alpha + s(\beta) & = s(\alpha + \beta) \\ \alpha + \beta & = \cup(\alpha + \lambda) : \lambda < \beta \text{ if } \beta \in \text{ Lim} \\ \alpha \cdot 0 & = 0 \\ \alpha \cdot s(\beta) & = s(\alpha \cdot \beta) \\ \alpha \cdot \beta & = \cup(\alpha \cdot \lambda) : \lambda < \beta \text{ if } \beta \in \text{ Lim} \end{array}$$

The operation $s(\alpha)$ can also be written as $\alpha + 1$. Note that addition and multiplication are associative but not commutative. For any natural n, in particular $n + \omega_0 = \omega_0 < \omega_0 + n$. Similarly, $2\omega_0 = \omega_0 < \omega_0 + \omega_0 = \omega_0 \cdot 2$.

Assuming the axiom of choice, the cardinal α of a set E can be defined

as the smallest von Neumann ordinal equinumerous to it, i.e., the cardinal is identified with its initial ordinal. Finite cardinals are likened to natural numbers. Zero is the cardinal of the empty set, $n > 0$ is the cardinal of $0, ..., n-1$. The cardinality of the set of naturals $\mathbb{N}$ is denoted $\aleph_0$. Indeed, $\aleph_0$ is also the cardinality of all countable infinite ordinals, as these ordinals are equinumerous to $\mathbb{N}$. $\aleph_0$ can thus be equally identified with $\omega_0, \omega_0^2, \omega_0^{\omega_0}$, etc. The arithmetic on ordinals differs from that on cardinals: although $\omega_0^2 > \omega_0$, we have $\aleph_0^2 = \aleph_0$. In this sense, a cardinal is a measure of an ensemble's "size", without regard to its order.

The infinite cardinal directly greater than $\aleph_0$ is denoted $\aleph_0^+$, or $\aleph_1$. This cardinal can be identified with ω_1, as well as with all ordinals resulting from a finite number of operations on ω_1. Note that if we accepted the continuum hypothesis, we would have $\aleph_1 = 2\aleph_0$, the cardinality of the set of reals $\mathbb{R}$. Then $\aleph_1^+ = \aleph_2$, then $\aleph_2^+ = \aleph_3$, etc. In general, every infinite cardinal is written $\aleph_\theta$, where θ is an ordinal number.

The cardinal $\aleph$ of a set E is said to be regular when it equals its own cofinality, meaning it equals the cardinality of the smallest set E' such that for all $a \in E$, there exists $b \in E'$ with $a \leq b$. Under the axiom of choice, α can equivalently be considered regular if for any index set I and for any sequence of sets $S_i (i \in I)$ such that $card(I) \leq \alpha$ and $card(S_i) < card(I)$, we have $card(\cup S_i : i \in I) < \alpha$. It is this second definition that will be most frequently used. Note that $\aleph_0$, as well as all successor cardinals, are regular. In contrast, $\aleph_\omega$ is singular.

A cardinal $\aleph_\theta$ is called a limit cardinal if θ is a limit ordinal. A cardinal $\aleph_\theta$ is called weakly inaccessible if it is a limit and regular. It is called strongly inaccessible if it also satisfies the condition $\lambda < \aleph \rightarrow 2^\lambda < \alpha$. All inaccessible cardinals are large cardinals, greater than $\aleph_\omega$, and constructed using smaller cardinals. Under ZFC axioms, it is not possible to prove the existence of an inaccessible cardinal. Indeed, ZFC theory is supposed to be consistent. However, adding the proposition "There is no inaccessible cardinal" to ZFC gives a theory that is also consistent. Consequently, the problem is undecidable. It should be noted that under the continuum hypothesis, notions of weak and strong inaccessibility merge.

Appendix 6: Shannon's Information Theory

> Information is the resolution of uncertainty.
>
> Claude Shannon

Information can be understood in several ways: from a semantic point of view, it is the intuitive and cognitive content of a statement, its meaning, and can be seen as what this statement teaches about the world. In this regard, tautologies such as "If he hadn't died, he would still be alive" teach nothing unknown *a priori* and therefore do not constitute information. Similarly, redundancies like "Napoleon's white horse is white" are poorer in information than innovative scientific propositions. Finally, the amount of information gained by knowing a particular event has occurred or will occur is greater the less probable the event. For instance, learning that, as per statistics, it rained across the Breton coast offers less information than discovering, against all odds, there was a heatwave.

From a syntactic point of view, however, information is the quantity of logical or linguistic symbols required to transcribe a statement in a given language. The higher this number, the denser the information. In this sense, tautologies are full-fledged information, increasingly significant as they require more characters to express; conversely, propositions rich in meaning, such as "$e^{i\pi} = -1$", can be succinctly formulated, thus being very poor in information from a syntactic viewpoint.

At first glance, these two approaches to what constitutes information seem difficult to reconcile and are, in many ways, mutually exclusive. Yet, it is legitimate to pose the following two questions: (1) Is it possible to formalise both in a mathematical language using quantifiable expressions that are comfortable to manipulate yet rich enough to capture intuition? (2) If so, can these two distinct expressions be generalised into a single, broader formula, under which the two previous definitions are merely special cases?

The statistical theory of information has answered affirmatively to both these questions, ultimately exhibiting a concept of informational quantity

known as entropy. First introduced by Shannon in his article *A Mathematical Theory of Communication* (1948), this theory successfully synthesises existing axiomatic approaches and pinpoints their primary areas of application.

The following recaps the foundations of information theory, presenting an axiomatic approach. Subsequently, the main measures of information and their properties are stated and methodically demonstrated.

Historical and Mathematical Perspective

Hartley and Wiener's Approaches

Hartley proposed in 1928 a definition of information quantity based on memory space, or the number of bits required to encode an arbitrary set of elements. For example, with 2^k elements, we know that k bits will suffice. By extension, let $\mathcal{X}$ be a set of arbitrary size. We write

$$H_{\text{Hartley}}(\mathcal{X}) = \log_2(|\mathcal{X}|).$$

If $\mathcal{X}$ is an alphabet, it can be used to write a word x. The information quantity contained in x corresponds to the bits required to encode the n letters of this word:

$$H_{\text{Hartley}}(x) = n \log_2(|\mathcal{X}|).$$

According to this view, information quantity only depends on the alphabet size, disregarding the meaning of words or the frequency of letters in this alphabet. However, smarter coding would readily exploit letter recurrences to reduce the required memory, assigning shorter codes to more frequent words and longer ones to less frequent ones.

This observation suggests to Wiener another definition of information quantity, based not on alphabet size but on the frequency of its letters. The underlying intuition is as follows. Suppose a reader aims to decode a message. Each letter deciphered provides information about the message, which is more valuable if the letter is rare in French. The measure H of the quantity of information acquired by reading a letter a depends on the frequency p of that letter in the alphabet. Specifically, $H(p)$ must be zero when $p = 1$ and tend towards infinity as p approaches zero. Between these extremes, $H(p)$ evolves monotonically. In 1948, Wiener constructed a function meeting these properties:

$$H_{\text{Wiener}}(p) = -\log_2(p).$$

Moreover, he proposed a set of axioms under which this function is uniquely defined. They are three in number and are as follows: given any

event of probability p and another of probability q, H_{Wiener} is the function such that:

(*i*) **Positivity** $H_{\text{Wiener}}(p) \geq 0$,

(*ii*) **Additivity** $H_{\text{Wiener}}(pq) = H_{\text{Wiener}}(p) + H_{\text{Wiener}}(q)$,

(*iii*) **Normalisation** $H_{\text{Wiener}}(\frac{1}{2}) = 1$.

The first axiom justifies that all reading provides information. At worst, this information might not be useful, but it cannot impoverish the reader's knowledge. The third axiom normalises, ensuring uniqueness. Axiom (*ii*), however, carries significant meaning and warrants discussion. Suppose a reader reads two letters a and b from an alphabet $\mathcal{X}$, with respective frequencies p and q. Let's call A the event of reading letter a and B the event of reading letter b. Further, suppose these readings are statistically independent events. In this condition, the probability of event $A \frown B$, or reading both letters in succession (in any order), is pq. Yet, the information provided by $A \frown B$ is the sum of information from A and information from B. We can thus write $H_{\text{Wiener}}(pq) = H_{\text{Wiener}}(p) + H_{\text{Wiener}}(q)$.

It is easily checked that under these axioms, H_{Wiener} indeed fits the monotonic decrease. Given two probabilities p_1 and p_2 such that $0 \leq p_1 \leq p_2 \leq 1$, write without loss of generality $p_1 = qp_2$, with $0 \leq q \leq 1$. We have:

$$\begin{aligned} H_{\text{Wiener}}(p_1) &= H_{\text{Wiener}}(qp_2) \\ &= H_{\text{Wiener}}(q) + H_{\text{Wiener}}(p_2) && \text{(by axiom ii)} \\ &\geq H_{\text{Wiener}}(p_2) && \text{(by axiom i)} \end{aligned}$$

Note the logarithmic base in entropy can be taken as $a \neq 2$. This requires altering the normality axiom to state $H_{\text{Wiener}}(1/a) = 1$ instead of $H_{\text{Wiener}}(1/2) = 1$. With the natural logarithm ($a = e$), the unit is the nat.

Shannon's Approach

Our question now concerns the average information a reader gains when discovering the first letter of a random word. Assume the letter is transmitted by an emitter E via a perfect (i.e. noiseless) channel and hence, no reading errors occur. In this case, the revealed letter is represented by a random variable X taking values in $\mathcal{X}$. Assume $m = |\mathcal{X}|$ and denote p_i the frequency of the i^{th} letter of the alphabet $\mathcal{X}$ (with $i = 1, ..., m$). The probability of reading an a depends on the frequency p_a of the letter and provides information equal to $H_{\text{Wiener}}(p_a)$. The probability of reading a b depends on the

frequency p_b of the letter b and provides information equal to $H_{\text{Wiener}}(p_b)$, etc... The information quantity linked to reading this random variable is thus given by the expected value over X, denoted E_X, of Wiener's entropy:

$$H(X) = E_X[H_{\text{Wiener}}(p_i)] = -\sum_{i=1}^{m} p_i \log(p_i).$$

This quantity is called Shannon entropy. It is more significant when the probabilities associated with each letter of the alphabet are comparable and reaches a maximum in equi-probability. If all symbols are equally frequent, indeed, the reader has no *a priori* clue about X's state, hence acquiring more information upon discovering it. Let's illustrate with an example.

Example 5. *Consider the random variable X as follows:*

$$X = \begin{cases} a & \text{with probability } p \\ b & \text{with probability } 1-p \end{cases}$$

If $p = 1$ or $p = 0$, we always have $X = a$ or $X = b$ (respectively). In this case, it is evident that reading the symbol provides no new information. Conversely, if $0 < p < 1$, we can deduce $X = a$ with probability p, but certainty is absent. The information provided by the sender resolves this uncertainty and offers information. Information thus relates to uncertainty. Greater uncertainty makes information more intriguing. The curve $H(p) = -p\log(p) - (1-p)\log(1-p)$ is depicted below.

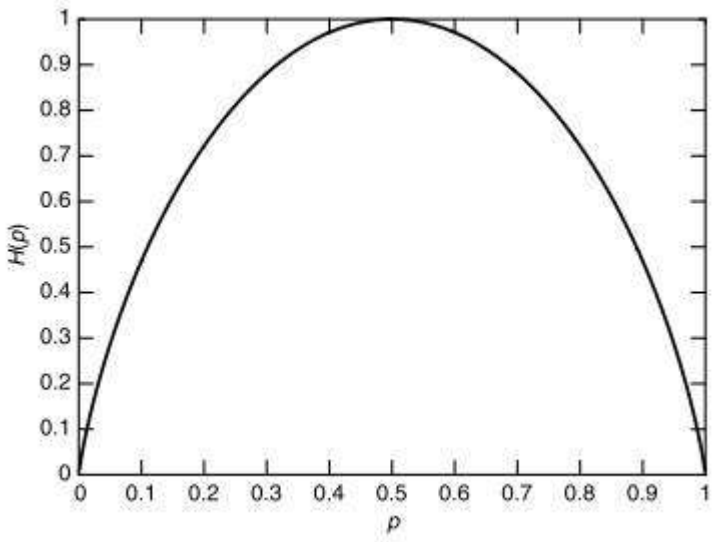


Figure 4: $H(p)$ as a function of p.

Entropy definition can also extend to event sequences where the sum of probabilities is less than 1. Let $p_1, ..., p_m \in [0, 1]$ where $\sum_{i=1}^{m} p_i \leq 1$. We write H_m as the m-argument entropy function defined as follows:

$$H_m(p_1, ..., p_m) = -\sum_{i=1}^{m} p_i \log(p_i).$$

Notice this expression generalises Hartley's and Wiener's entropy. In the case where $p_1 = ... = p_m = 1/m$, we indeed have $H(X) = \sum_{i=1}^{m} 1/m \log(m) = \log(m) = H_{\text{Hartley}}(\mathcal{X})$. When $m = 1$, $H_1(p_1) = -\log(p_1) = H_{\text{Wiener}}$.

Axiomatic System

Under certain axioms, Shannon entropy can be uniquely derived. The axiom selection varies in literature [**?**], [**?**] but all converge to the same definition of $H(X)$. The axioms presented here are those chosen by Shannon himself.

Shannon's Axioms

- **Normalisation:** $H_2(\frac{1}{2}, \frac{1}{2}) = 1$.
- **Continuity:** For any $p \in [0, 1]$, $H_2(p, 1-p)$ is a continuous function of p.
- **Grouping:** For any natural number m and any sequence of reals $p_1, ..., p_m$ such that $\sum_{i=1}^{m} p_i \leq 1$: $H_m(p_1, ..., p_m) = H_{m-1}(p_1 + p_2, p_3, ..., p_m) + (p_1 + p_2)H_2(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$.

It is possible to show that the unique function H_m satisfying these three axioms writes:

$$H_m = -\sum_{i=1}^{m} p_i \log(p_i).$$

Due to the length and complexity of this proof, we invite readers to refer to the appendices. Here, we only highlight and comment on the key steps of the approach. The core idea involves characterising the function $H_m(\frac{1}{m}, ..., \frac{1}{m})$, which corresponds to a uniformly distributed random variable's entropy, showing from the axioms it equals $\log(m)$. This approach parallels Wiener's, differing only in the initial axiom set. Achieving this necessitates demonstrating and using the following lemma:

Theorem 12. *If the function $f(m)$ satisfies:*

- $f(mn) = f(m) + f(n)$
- $\lim_{n\to\infty}(f(n+1) - f(n)) = 0$.

- $f(2) = 1$

then $f(m) = \log_2(m)$.

Once established, we can show for any rational p, $H_2(p, 1-p) = -p\log(p) - (1-p)\log(1-p)$. Using continuity, this result can extend to irrationals. It is finally generalized to all $m \geq 2$ by an induction process utilising the grouping axiom (in a generalised form).

Note that adding the monotonicity axiom, the lemma proof above becomes substantially simpler. This axiom states that for all $m \in \mathbb{Z}^+$, $f(m) = H_m(\frac{1}{m}, ..., \frac{1}{m})$ is a non-decreasing and monotone function of m.

Numerous alternatives to the proposed axiomatic approach exist. Almost every entropy property has been axiomatically demonstrated to prove the entropy formula. The challenge is choosing a minimal set number of properties while ensuring they are intuitively sound.

Shannon's Measures

The following section summarises different entropic magnitudes: joint entropy, conditional entropy, Kullback-Leibler divergence, mutual information, and so on. We first outline all definitions, then prove essential inequalities involving these measures. The definitions and proofs are conducted in multivariate form.

Joint and Conditional Entropies

For any $n \geq 2$, consider discrete random variables $X_1, ..., X_n$ taking values in $\mathcal{X}1, ..., \mathcal{X}n$ (respectively). These n random variables have $k = 2^n - 1$ joint probability functions. Let $\mathcal{N} = 1, ..., n$. For any $\alpha = (i_1, ..., i_m) \subseteq \mathcal{N}$, denote $p\alpha$ the joint probability function for the variables $Xi_1, ..., X_{i_m}$. Following the same logic, denote $X_\alpha = X_i : i \in \alpha$, $\mathcal{X}_\alpha = \mathcal{X}_i : i \in \alpha$. The notion of entropy extends to a set of more than two random variables.

Joint Entropy

Let $\alpha \subseteq \mathcal{N}$. *Joint entropy* of variables X_α is written

$$H(X_\alpha) = H(\alpha) = -\sum_{x \in \mathcal{X}\alpha} p\alpha(x)\log(p_\alpha(x)).$$

By convention, assume $H(\emptyset) = 0$.

Similarly to conditional probability, conditional entropy is also recognised. Defined by replacing probabilities with conditional probabilities in the entropy definition, it is:

Conditional Entropy

For $\alpha, \gamma \subseteq \mathcal{N}$. *Conditional entropy* of variables X_α with respect to X_γ is written:

$$H(X_\alpha|X_\gamma) = H(\alpha|\gamma) = -\sum_{x\in\mathcal{X}\alpha} \sum y \in \mathcal{X}\gamma p\alpha(x|y)\log(p_\alpha(x|y)).$$

By convention, write $H(\alpha|\beta) = H(\alpha)$ when $\beta = \emptyset$.

Relative Entropy

Relative entropy, also known as Kullback-Leibler divergence, defines a pseudo-distance in the space of statistical distributions. It aims to quantify the difference between two probability densities p and q. In particular, it is zero when p and q are identical on supp(p), the compact support of p. Otherwise, it is always positive. Relative entropy reveals nothing about p and q outside supp(p), which is its primary limitation.

Relative Entropy

Let two probability densities p and q be defined over the same domain $\mathcal{X}$. Relative entropy (or Kullback-Leibler divergence) between p and q, denoted $D(p||q)$, is defined as:

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)f(x),$$

where f is the continuous extension of $\log\frac{p(x)}{q(x)}$

$$f(x) = \begin{cases} 0 & \text{if } p(x) = 0, \\ \infty & \text{if } p(x) \neq 0 \text{ and } q(x) = 0, \\ \log(\frac{p(x)}{q(x)}) & \text{otherwise.} \end{cases}$$

Note that if some $x \in \mathcal{X}_\alpha$ such that $q(x) = 0$ and $p(x) \neq 0$, relative entropy is unbounded.

Relative entropy fulfils two-thirds of metric axioms. It is a pseudo-distance, not strictly a distance, as the Minkowski inequality is not satisfied. Also, $D(p||q)$ is not generally symmetric. For instance, if $x \in \mathcal{X}$ with $p(x) = 0$ and $q \neq 0$, $D(p||q) = 0$ while $D(q||p)$ is unbounded.

$$D(p||q) = 0 \Leftrightarrow p(x) = q(x) \qquad \forall x \in \text{supp}(p), \tag{1}$$

$$D(p||q) \geq 0. \tag{2}$$

The notation supp(p) refers to p's compact support. The proof of these properties is provided in the appendices.

Mutual Information

Another important quantity in information theory is mutual information. Defined based on entropy, it measures the information held about a set of random variables when another set of statistically dependent random variables is known. This is, in other words, a measure of the uncertainty reduction gained from knowing one set of random variables because of another.

Mutual Information

Let two sets $\alpha, \beta \subseteq \mathcal{N}$. Denote $p_{\alpha\cup\beta}$ the probability function corresponding to the union of variable sets X_α and X_β. *Mutual information*, noted $I(X_\alpha; X_\beta)$, is

$$I(\alpha;\beta) = D(p_{\alpha\cup\beta}||p_\alpha p_\beta).$$

For any set $\gamma \subseteq \mathcal{N}$, define *Conditional mutual information*:

$$I(\alpha;\beta|\gamma) = D(p_{\alpha\cup\beta|\gamma}||p_{\alpha|\gamma}p_{\beta|\gamma}).$$

By convention, we write $I(\alpha;\beta|\gamma) = I(\alpha;\beta)$ when $\gamma = \emptyset$.

If variable set X_α is statistically independent from X_β, for all $(x, y) \in \mathcal{X}\alpha\cup\mathcal{X}\beta$, $p_{\alpha\cup\beta}(x, y) = p_\alpha(x)p_\beta(y))$ and $D(p_{\alpha\cup\beta}||p_\alpha p_\beta)$ is zero. The converse doesn't hold. Suppose $p_{\alpha\cup\beta} = p_\alpha p_\beta$ throughout $p_{\alpha\cup\beta}$'s compact support but some point (x, y) not in it where $p_\alpha p_\beta$ isn't zero. Here, $D(p_{\alpha\cup\beta}||p_\alpha p_\beta) = 0$. Yet, X_α and X_β aren't statistically independent. With this restriction, mutual information is seen as a measure of statistical dependence. It is never negative (since relative entropy isn't).

Appendix 7: Essay - Is Consciousness Dead?

This thesis lays the groundwork for a mereological theory of consciousness but leaves numerous questions unanswered. Firstly, it is appropriate to ponder how the notions of the Self, the I, and the Ego, which I have merely sketched out, link with the three cognitive levels of Sensation, Perception, and Reason. Moreover, it seems essential to justify introducing the principle of differentiation, which I stated lies at the heart of the consciousness mechanism.

To elucidate these queries and give my argument more substance, I have written a 300-page literary essay entitled: "Is Consciousness Dead?" In its preliminary form, this text won the first prize in the Art and Letters of France competition in 2022, in the essay category. Despite this accolade, I have continually refined it to the shape presented today. Due to its poetic style, the personal and occasionally controversial tone adopted, and the liberties I have taken in idea development, I deemed it preferable not to include it in the body of this thesis. It is not a classic academic discourse debating existing theories to position within an expert debate. Instead, it is a personal stance, humanist and universalist, addressing a question touching the very condition of human existence: what am I as a conscious being? What still gives value to today's world?

The reader will find a brief presentation of the work attached. To better understand its framework, I also invite them to refer to the table of contents of the actual text.

Presentation

During a journey through northwestern Thailand, I had a strange encounter. The air was hot and humid, the sun hung vertically overhead, bathing the dusty paths in a blood-red halo of light. Sweat trickled down my body, and the mosquitoes reveled in their feast. It was a few months before the

start of my doctorate. Little did I know that the hardships endured in that inhospitable jungle were merely the prelude to a long journey through the world of ideas.

Despite the oppressive heat, it did not take long before I arrived at the outskirts of a monastery. As I stepped into the inner courtyard, I heard voices nearby. A young man was speaking with great eloquence, gesturing with every word—unless, perhaps, it was merely to swat away the mosquitoes? Beside him, an older monk listened in silence. As I drew closer, I realized that their conversation touched upon a subject close to my thesis. "Given the knowledge we have accumulated about the brain," the young man declared with passionate conviction, "we are now allowed to think that consciousness, as we experience it, might be an illusion. Do you know what struck me the most when I began studying neurology? The realization that all my sensations are nothing but electrical currents." Opposite him, the monk simply smiled. He interjected with occasional questions but never offered his own opinion. When their conversation came to an end, he suggested they go eat. But just as he was about to stand, he hesitated: "Will you," he asked, "be able to endure the illusion of food?"

This remark made me smile, for it seemed to illustrate quite well a tension that has existed since the dawn of the modern era between science and personal experience. From a purely introspective and intimate standpoint, the only undeniable certainty for a Buddhist monk is that, at any given moment, he is conscious of something rather than nothing. The tangible world that unfolds beneath the exploration of his fingers, his eyes, his tongue, exists only in the consciousness he has of it. The very notion of being a subject perceiving the world, a body within matter, a thinking personality endowed with history, memory, and an immutable character, seems to be something that can be questioned. Only the present moment, in its fullness, truly exists.

From the perspective of scientific experience, however, it would seem that the primary certainty is the existence of a material world—composed of hadrons, waves, and energy—within which it is difficult to explain how the phenomenon of consciousness itself emerges. When I gaze upon the eternal sky, when I picture, one by one, the infinity of worlds scattered like stars across the vast canvas of the Milky Way, each one lost in the frozen abyss of the void, I cannot help but feel a strange melancholy. The universe appears to me terrifyingly more real than the consciousness I have of it. It is populated with sounds, colours, intoxicating fragrances, which are nothing more than quantum particles in stochastic motion. What I take to be an authentic phenomenon of consciousness—the contemplation of the sky—is nothing but a bit of sound and fury, a cascade of neural reactions in the brain.

Throughout this essay, I will attempt to prune the great tree of Western metaphysics—a trunk blackened by years of philosophical battles—in order

to free it from the grip of these two prejudices. The first consists in saying (1) that consciousness is within the brain. The second, conversely, asserts (2) that the world is entirely within the consciousness we have of it. These two suppositions are as old as materialism and idealism. They have opposed one another since time immemorial, wielded by philosophers in their dressing gowns, yet perhaps they have never been so tenacious? And for good reason: a single idea is nestled within both positions. This idea, as I will endeavour to show in this essay, is that consciousness is something individual.

We take it as self-evident that consciousness is a personal fact; all the notions, sensations, and perceptions that pass through us belong to us alone. The idea that our minds might contain experiences that do not belong to us would strike us as strange. Yet this was the belief of the Ashuar people in the Amazon, who saw dreams as a space where their thoughts mingled with those of plants and animals. It was also commonly accepted within the Christian faith that a believer could hear the word of God within themselves. It was therefore not uncommon to conceive of consciousness as something that exceeded the bounds of the individual mind; human mental life was interwoven with that of God (Christianity), Brahma (Hinduism), Spirit (Spinoza), the Absolute (Hegel), the One (Plotinus), or Nous (Anaxagoras).

Today, things are quite different. Anyone who claims to receive transmissions of thought, to feel the presence of the dead within them, or to communicate with a higher consciousness—whether nature, God, or a spirit—is suspected of having a neurological disorder. Ever since consciousness has been regarded as an individual fact, existing monadically and independently of any other consciousness, philosophy has been left with only two metaphysical possibilities: either consciousness is located somewhere within the individual's body (the stance of reductionist materialism), or the entire world is contained within the consciousness of that individual (the stance of solipsistic idealism). This methodological excluded middle prevents us from considering that consciousness might extend beyond the boundaries of our body (the position of extended cognition) or that it might exceed the limits of the self (the position of non-solipsistic idealism).

In this essay, I will argue that both of these positions—the one that treats consciousness as a property of the brain (reductionist materialism) and the one that treats the material world as a property of consciousness (solipsistic idealism)—lead to contradictions. They both fail to distinguish, in the study of the human mind, two very different notions:

- S**ubjective Consciousness:** On the one hand, consciousness refers to a state—the subjective fact of being conscious, here and now, of something rather than nothing. This is not the awareness of any particular thing—the table before me, the wind in my hair, etc.—but rather the sheer faculty of being conscious of potentially anything at

all. As long as we possess it, this faculty is what allows us to say that we are alive. For this reason, I will also refer to it as living consciousness.

- **Objective Consciousness:** On the other hand, consciousness also designates an object in the world, endowed with physical or psychological properties: the mind, the self, the soul, the brain, etc. This object—just as concrete as the representation of a table before me or the sensation of wind in my hair—is presumed to be the cause of all possible conscious experience. Yet this objective consciousness is not the totality of mental processes. It is neither an organic phenomenon nor a faculty. It is a dead consciousness, awaiting, like all other observable objects, to be dissected by the scalpel of scientific method.

In reductionist materialism, objective consciousness takes the form of a brain equipped with neurons, synapses, and gray matter, whose exuberant activity is thought to produce thought itself. In solipsistic idealism, objective consciousness is treated as a kind of passive receptacle, a self that is the source of the mental realm—one to which all lived experiences could be attributed. In both cases, an ingenious sleight of hand is at play: subjective consciousness—which is nothing but a simple here and now—is given an objective, spatial, and temporal form.

This conceptual conjuring trick runs up against a logical paradox, which can be illustrated through an instructive analogy. Suppose an individual has spent their entire life in a room devoid of any reflective surfaces. They can see the world around them—the walls, tables, and chairs. They can also look at their hands, their body, and—with a bit of luck—the tip of their nose, reddened by too much alcohol. But they have never seen their own face or the colour of their eyes. How could such a person explain the phenomenon of vision?

Our hypothetical individual would certainly understand that vision, as a daily experience, must have its origin within him. He might even assert that this origin lies somewhere on his face. Yet, though he can close his eyelids and feel his eyes with his fingertips, he would not be able to see his own eyes with his own eyes. In other words, vision would be for him a sensory certainty but not a visual one. He would be aware of his own eye as something abstract—something he can touch with his hands, something he can feel within himself, but something he cannot see.

What happens to this analogy when applied not to the eye of vision but to consciousness? Just as the eye itself does not fall within the scope of sight, it may be that consciousness, as an experience, cannot be grasped through the faculties of consciousness itself. Each time I conceive of myself

as consciousness—each time I tell myself, "I am this person," "I am the one who thinks," "I am cogito," "I am the subject of all experience," or "I am a brain in the world"—each time I engage in self-contemplation, I find before me an image, an emotion, a sensation of myself. This is an object of my thought like any other and is no different, in this respect, from my representation of a table, a book, or a monkey. This image of myself is not my objective consciousness as such, but merely a subjective representation contained within my conscious experience.

Perhaps this apparent paradox is the fundamental reason why the world of consciousness seems populated with objects that exist within me but are not me? The central thesis of my essay will be to argue that, unlike ordinary objects, consciousness cannot be represented. It can either be lived subjectively as an "I" (I am consciousness), or conceived abstractly and objectively as an "It" (there is consciousness). In the first case, consciousness is an ineffable something about which nothing can be said. It manifests itself, but the moment it appears, it is already gone. In the second case, it is nothing more than an abstract idea—something that cannot be concretely represented. This abstract consciousness is somewhat like the proposition: "Homo sapiens is the being that thinks that it thinks." We understand that Homo sapiens thinks about the fact that it thinks, but we do not know what it is thinking about concretely. Thinking about the fact of thinking does not seem as vivid as thinking about the tempting apple in the middle of my garden.

Given the perilous dead ends looming over its head, one might be tempted to discredit consciousness altogether. Indeed, some scientists claim that consciousness is an illusion—that outside this mass of cells, nerves, and tissues—this assemblage of dead matter—there is nothing truly subjective. Seeking the presence of conscious life within the brain would be akin to a fruitless treasure hunt. I agree with this view, but only if we accept consciousness as an individual reality. An individual, in essence, is defined as an indivisible part of a greater whole—a citizen within society, a living being within the animal world, a collection of atoms within the universe—whereas consciousness, as a subjective experience, is both everywhere and nowhere at once. Everywhere, because everything we see, think, and experience is a subjective phenomenon of consciousness. Nowhere, because nothing we see, think, or experience is objective consciousness itself. This is why, when we seek consciousness in matter, we find it nowhere; but when we seek consciousness within conscious experience itself, we find it everywhere. It follows that consciousness cannot be conceived as a mere part of the whole. Rather, it stands, in Wittgenstein's words, at the border of the whole. It is the very limit of our individuality.

I do not deny that there is something deeply individual about conscious experience. This individual aspect is the self—the "I" that thinks, feels, and experiences. Indeed, the recognition of oneself as an autonomous "I" worthy

of consideration is among the most fundamental and precious rights. But this self is not consciousness in its entirety—it is only a fragment of it. The self, in fact, exists only in opposition to a non-self. When I say "I am wise," "I am healthy," or "I am young," I simultaneously reject thoughts that are unwise, sensations that seem sickly, and impressions of aging. In Freudian terms, the self is a construct formed by repressing a part of itself. Even when we gaze into a mirror, point at the reflection of our body, and exclaim, "I am this body," we are performing an act of separation. We isolate one object of consciousness—the perception of this body—from all other existing perceptions: those of the mirror, the room, and the objects within it. Yet, in a certain sense, the room, the objects, and the mirror are also within my consciousness. They exist within me, but they are not the self as such.

Of course, the self is never definitively acquired. Rather, we seem to shift through different representations of the self throughout our psychological development. These facets of ourselves—these personas—may come into conflict, contradicting, affirming, or negating one another in complex relational patterns. "I is another," as the poet said. This is why the most pressing task of contemporary metaphysics may be to distinguish the self from the I. The self is the ceaselessly renewed attempt to objectify oneself as an individual. The I, on the other hand, is the certainty of being all things—that nothing exists except consciousness. The I has a universal quality; it is both within us and beyond us, dissolving the boundaries between subject and object. The self, in contrast, remains individual.

If we understand things in this way, the central problem becomes how exactly the I (subjective consciousness) and the self (objective consciousness) relate to one another. The task of this essay will be to determine the dynamic principles that allow consciousness, paradoxically, to conceive of itself as an individuality. In many respects, these principles resemble a struggle. The self is like the desert wanderer who piles sandbags in the scorching dunes, who erects makeshift barriers, who builds windbreaks of cloth to shield himself from the storm—before crying out into the howling winds: I am, I am alive! Perhaps he would do better to reinforce his walls? To open a gap in the fabric of his tent? To let the wind pass through? Perhaps then he would hear the wind whisper softly: I am life, I am invisible, I am living consciousness. For consciousness is not dead—far from it!—it has merely given shape to Maeterlinck's words:

> *It is rare for a mystery to vanish; more often, it merely changes place.*