UCLouvain

UCLouvain — cog-SUP

Human-Like Game-Playing Agents

Beyond Behavioural Alignment

Author: **Aloïs Rautureau**

Supervisor: **Éric Piette**

Track: **Computational Neuroscience and Artificial Intelligence**

The chess literature is for the most part of a purely technical nature. It deals with the play and not with the player and their way of thinking; it treats the problem and not the problem solver.

Adriaan D. de Groot

Declaration of originality

The research project of this thesis consists in exploratory work to broaden and define human-like game-playing agents, derive methods to quantify behavioural alignment, and study search algorithms and heuristics for games inspired by human cognition. Previous research has studied human-like agents primarily through the lens of behavioral alignment, although some work has been done on cognitively-inspired methods before the widespread use of neural networks. As of today, work on human-like game-playing agents lacks a cohesive or comprehensive definition of their object of study, a gap which we aim to fill during the first half of this thesis through the study of psychological and anthropological insights. The second half will focus on exploring the potential of cognitively-inspired methods, which are broadly outdated, from a modern point of view.

Declaration of contribution

- Defining the scientific question: Aloïs Rautureau, Éric Piette
- Bibliographic research: Aloïs Rautureau
- Scientific methodology: Aloïs Rautureau
- Coding, data collection and analysis: Aloïs Rautureau
- Writing the dissertation: Aloïs Rautureau
- Proofreading and feedback: Éric Piette
- Co-authors on scientific papers related to this dissertation: Aloïs Rautureau, Éric Piette, Tristan Cazenave, Loris Sogliuzzo, Tim Penn, Summer Courts, Walter Crist, James Goodman, Dennis J.N.J Soemers, Alex De Voogt, Hendrik Baier, Jacob Schmidt-Madsen (Rautureau, Cazenave, & Piette, 2026; Rautureau, Piette, Penn, et al., 2026; Sogliuzzo et al., 2026)
- Figures: Aloïs Rautureau (unless stated otherwise)

Abstract

Research on game-playing artificial intelligence has historically treated games as benchmarks of intelligence. With artificial agents surpassing human players in many games over the years, this original goal has shifted towards designing agents that are aligned with human behavior to provide more engaging experiences, actionable strategies, or simply to study human playing styles. Modern approaches to "human-like" play mainly focus on behavioural alignment, broadly defining human play from a functionalist viewpoint derived from evaluation frameworks similar to the Turing test. Although significant advances have been made through these lenses, we argue that it narrows the definition of play, overlooking social, interactive, and cognitive dimensions. While these aspects formed an active field of research in the past—mainly with the goal of approaching the strong intuition of human experts—they have become secondary in modern works. Beyond their interest to the field of artificial intelligence, a broader definition and study of human-like agents aims to support richer computational models to investigate the cognitive processes involved in decision-making, and the cultural and historical aspects of games.

This thesis aims to reframe the definition of "human-like" agents by studying game-playing not only as a sequence of moves to reach a rule-defined goal, but as an inherently social, culturally-situated, and cogntitively-grounded activity. To this aim, we propose a multi-dimensional classification of human-like game-playing agents, review current methods for measuring behavioral alignment, and explore the prospect of cognitively plausible search algorithms in games. Beyond these exploratory assessments, we discuss future work and research directions for a broader study of human-like play.

Acknowledgments

This thesis was written with the help of numerous and thoughtful comments from my supervisor, Éric Piette, and would not be as complete without the invaluable insights and encouragements from fellow GameTable COST Action members on aspects related to psychology, sociology and game-playing systems. I'd especially like to thank my co-authors on the vision paper "Human-Like Game AI Beyond Behavioural Alignment", submitted during the writing of this thesis, which was foundational to many arguments that made their way into its chapters: Tim Penn, Summer Courts, Walter Crist, James Goodman, Dennis J.N.J Soemers, Alex De Voogt, Jacob Schmidt-Madsen, Hendrik Baier and Graham Todd.

I also extend my gratitude to Tristan Cazenave for co-supervising the first semester of my internship, centered around cognitively plausible search algorithms.

Thanks to UCLouvain, Schmidt Science and GameTable COST Action for financially supporting various aspects of this work.

Finally, I want to thank my family, friends and significant other for their unyielding support throughout the many hardships, joys and surprises that come from writing a master's thesis.

The contents of this thesis have been the source of multiple published or submitted works:

- *Generalized Rapid Action Value Estimation in Memory-Constrained Environments*, A. Rautureau, T. Cazenave, E. Piette, published at the ICGA Computers and Games 2026 conference (Rautureau, Cazenave, & Piette, 2026)
- *Game AI for Cultural Heritage: Outcomes from the GameTable 3rd Grant Period Opening Meeting*, A. Rautureau et al., published in the ICGA Journal (Rautureau, Piette, Penn, et al., 2026)
- *Human-Like Game AI Beyond Behavioural Alignment*, A. Rautureau et al., in proceedings of the IEEE Conference on Games 2026 (*submitted*) (Rautureau, Piette, Goodman, et al., 2026)
- *Toward modelling player-specific Chess behaviors*, L. Sogliuzzo, A. Rautureau, E. Piette, in proceedings of the IEEE Conference on Games 2026 (*submitted*) (Sogliuzzo et al., 2026)

Table of Contents

1

Introduction

The study of game-playing systems has historically been focused on designing artificial agents capable of playing as effectively as possible (Yannakakis & Togelius, 2025). In practice, this translates to optimizing for **winning**. Games are of peculiar interest to the field of Artificial Intelligence (AI) due to their inherently symbolic nature, allowing the implementation of highly efficient search algorithms while being sufficiently complex environments to require the use of advanced heuristics. As of writing this thesis, the state-of-the-art game-playing agents all combine a tree search algorithm (AlphaBeta (Knuth & Moore, 1975) or Monte-Carlo Tree Search variants (Kocsis & Szepesvári, 2006)) and deep neural networks to guide said search, achieving superhuman playing strength in many games, with Chess (Silver, Hubert, et al., 2017), Go (Silver, Schrittwieser, et al., 2017) and video games like StarCraft II (Vinyals et al., 2019) or DotA 2 (Berner et al., 2019) being prime examples.

In contrast, research in designing artificial agents that play *like humans* has had a secondary role. It is important to make the distinction between *human-like* and *human-level* agents in the context of game-playing early on. Artificial agents are fully capable of playing at a human-level through placing arbitraty limitations on their computational capacities, but the resulting playing patterns have been anecdotally observed to be synthetic and unnatural.

While a form of cognitive mimetism has been involved in developing strong heuristics —an area in which humans empirically surpassed machines until the advent of deep learning and neural networks, the main goal remained enhancing and surpassing human capabilities. As the importance of Human-AI Interaction grows by the day, this area of research has become active once again.

1.1. Motivating human-like agents

Superhuman playing strength in games, while a feat in its own right, yields little benefit to players aiming to learn clever strategies, face believable opponents or cooperate with these programs in multiplayer games. In collaborative tasks for example, agents that modelled human behavior as part of their decision process were shown to perform better than if they modelled their collaborators as optimal agents (Carroll et al., 2019; Turnwald & Wollherr,

2019). Having access to human-like agents could also improve the reliability of data mining for game-related metrics like drama, average game length or knowledge transferability.

Behavioral alignment—a quality of artificial agents to mimic the playing style of human players—is certainly a first and important step towards closing that gap. However, it still focuses on the symbolic nature of games rather than trying to understand players themselves. An agent mimicking a human playstyle may provide insights about decisions taken, but not *how* or *why* they were taken. More generally, the study of games from the perspective of AI research has a tendency to reduce them to puzzles that need to be solved, eclipsing the broader social and interactive nature of game-playing. We argue that this narrowing of play has carried over to modern studies of human-like game-playing. For example, when attempting to teach strategic concepts through analysis of positions, decisions are not sufficient even if the model is explainable: they need to be actionable by human players in real gameplay, which requires taking into account cognitive constraints on working memory and rationality.

1.2. Structure and contributions

Research on human-like agents appears fragmented: definitional issues are often tackled in passing, leaning mainly on the Turing test (Turing, 1950), and measurement methods to assess the human-likeness of produced models lack extensive study and psychological grounding. This has two consequences: first, it tends to narrow game-playing to a sequence of individual, unrelated decisions, eclipsing the material, interactive and cognitive nature of play; second, the lack of cohesion in metrics used to measure "human-likeness" makes it difficult to compare models. This thesis aims to unify this field in two ways: providing cohesive classifications and metrics for human-like agents, grounded in studies of human game-playing; and broadening potential research directions by focusing more intently on the mechanisms from which human gameplay arises, both internal and external.

Introductory and select reminders on game theory and game-playing agents, which are the main focus of this manuscript, can be found in Chapter 2. Chapter 3 provides a deeper dive into studies on the social and cognitive aspects of human play.

Building upon this foundation, Chapter 4 provides a discussion on how to define "human-likeness" in the context of game-playing. It contains the first contribution of this thesis as a classification of human-like agents centered around three axes: their ability to transfer knowledge across domains, their qualities as behaviorally aligned and/or cognitively plausible systems, and the granularity of this modelling with regard to the variety of human playing styles and motivations. Chapter 5 then provides a review of metrics measuring behavioral alignment, as well as our own contribution in the form of a statistical test acknowledging individual players rather than an averaged population. This chapter also discusses future plans towards quantifying the transferability of knowledge between games.

Chapter 6 takes a first step towards exploring cognitively plausible search algorithms for game-playing agents, an aspect which has been less extensively studied. This chapter is comprised mainly of novel contributions.

Finally, we conclude this thesis in Chapter 7 by discussing its contents more generally, as well as the many future research directions stemming from this work.

2

Computers playing games

Artificial game-playing agents have a long history of research, with games often being considered the "*Drosophila of artificial intelligence*" due to their nature as complex, yet constrained environments. As such, there is extensive literature on the variety of methods and algorithms that can be used to make computers play games (Yannakakis & Togelius, 2025).

Game-playing programs can be roughly split into two core components: (1) a game representation encoding the rules of the game, allowing to at least query the legal moves in a given state, make moves and check whether the game has ended; (2) utilities related to planning.

While (1) is not directly relevant to this thesis, it is necessary to establish general frameworks by which games are represented. (2) is directly responsible for dictating the behavior of the game-playing agent and requires an introduction of state-of-the-art techniques and formalisms to understand why current systems fail to capture human-like decision making.

Planning utilities in game-playing agents can be broken down further into search and heuristics. Search algorithms are formulated as logical inference frameworks allowing agents to emulate forward-planning by traversing a tree of states and propagating observations. Due to the complexity of most games, visiting their tree in its integrity is often intractable. Heuristic functions allow agents to incorporate long-term rewards by providing estimates about the value of states and moves. Note that this does not fully encompass the variety of methods used to program computers to play games. Some models are fully able to play with only one of these components, and some use neither.

This chapter begins by introducing the formalism used throughout this thesis to represent games Section 2.1. This naturally leads to manipulations of said representation by search algorithms, which we discuss in Section 2.2. Finally, Section 2.3 introduces the main forms of heuristic functions used in state-of-the-art agents.

2.1. Games as symbolic environments

While there are many ways to represent games—e.g. payoff matrices, graphs or their extensive form—this thesis will use Markov Decision Processes (MDP) (Vaserstein, 1969).

A MDP is described as a 4-tuple (S, A, P, R), with:

- S the set of **states**. This maps directly to the state of the board in tabletop games, the hands of players and the deck in card games, etc.
- A is the set of **actions** that an agent can take to possibly change its current state. In the context of games, we denote $A_s \subseteq A$ the actions an agent is allowed to take in a given state $s \in S$
- P is the **transition function**. Given a state-action pair $(s, a) \in S \times A$, it returns a probability measure over S indicating the probability for an agent to reach state s' by taking action a in state s. In perfect information games, there is only one state that can be reached per state-action pair. We will ignore the case of imperfect information or stochastic games, as they are outside the scope of this thesis.
- R is the **reward function**. It associates each triplet $(s, a, s') \in S \times A \times S$ with a corresponding value $r \in R$. For games like Chess or Go, it is often conceptually easier to think of rewards as win-loss scores obtained when reaching a terminal position s by playing move a.

MDPs model a graph of states linked together by actions. The trajectory of an agent within this graph can be written as a sequence of states and actions (1).

$$s_0 \overset{a_0}{\rightarrow} s_1 \overset{a_1}{\rightarrow} \dots \overset{a_{n-2}}{\rightarrow} s_{n-1} \overset{a_{n-1}}{\rightarrow} s_n \tag{1}$$

where $s \overset{a}{\rightarrow} s'$ denotes that the agent transitioned from s to s' by making action a.

An agent's behavior is modelled through its **policy** π. This policy associates each state $s \in S$ with a probability distribution over actions in A. $\pi(s, a) \in [0, 1]$ denotes the probability that an agent following policy π takes action a in state s. We will also use the notation $\pi(s)$ for the probability mass function over A for state s. By convention, $\forall a \notin A_s, \pi(s, a) = 0$.

MDPs were chosen because of their natural mapping to game-playing concepts, but also their generalizability to other domains. The MDP framework is central to fields like Reinforcement Learning (RL) (Sutton & Barto, 2018) due to their ability to model decision tasks within complex environments. By using this formalism, the methods discussed in this thesis might be applicable to many other domains interested in the development of human-like agents.

2.2. Search algorithms

Search algorithms work on the graph of states derived from the rules of the game as described in the previous section. They start from a root state—from which the move will be played—then explore the game graph to assess the value of each playable move. An important aspect of search algorithms is the way they backtrack and keep count of scores obtained in terminal nodes: while the agent wants to maximize its score, it is often assumed that their opponent will want to maximize their own. In zero-sum games—where the rewards obtained by players sum up to zero—this is equivalent to the opponent trying to minimize our score.

State-of-the-art search algorithms for games can be split in two categories. Depth-first algorithms (Minimax (Von Neumann, 1959), AlphaBeta (Knuth & Moore, 1975)) treat one branch at a time, playing a fixed number of moves before evaluating the resulting state, then backtracking. Best-first algorithms (Monte-Carlo Tree Search (Kocsis & Szepesvári, 2006), Proof Number Search (Allis, 1994)) keep statistics about visited states to try and

focus on the most promising branches according to a given heuristic, expanding an asymmetric search tree that is kept in memory.

This section specifically introduces Monte-Carlo Tree Search (MCTS) family of algorithms (Kocsis & Szepesvári, 2006), as they are the focus of subsequent chapters.

MCTS is a search framework rather than a single algorithm. It approaches the identification of the best move as a sequence of Multi-Armed Bandit (MAB) problems: given k arms with unknown rewards, the agent must identify the arm with the highest expected reward. In games, this maps to choosing the action that has the highest chance of leading to a winning terminal state. MCTS maintains a search tree in memory with associated expected rewards for each node. It updates these estimates by proceeding in four-step iterations (Figure 1): it first *selects* a leaf node, then *expands* it, *simulates* a possible continuation in order to obtain a reward signal, which is then *backpropagated* up the tree to update value estimates. The algorithm repeatedly applies these steps until it runs out of a predetermined budget—time, number of iterations and memory are the most common.

MCTS algorithms are derived from this framework by specifying strategies for each step. The most common selection strategy, Upper Confidence bounds applied to Trees (UCT) (Kocsis & Szepesvári, 2006), weighs both the current estimate of nodes (exploitation), as well as the number of samples that make up that estimate (exploration). It may favor nodes that have been sampled less over a node that has proven valuable if they have been less frequently visited. UCT selects the child n of node p that maximizes $\mathrm{UCT}(p, n)$ (2).

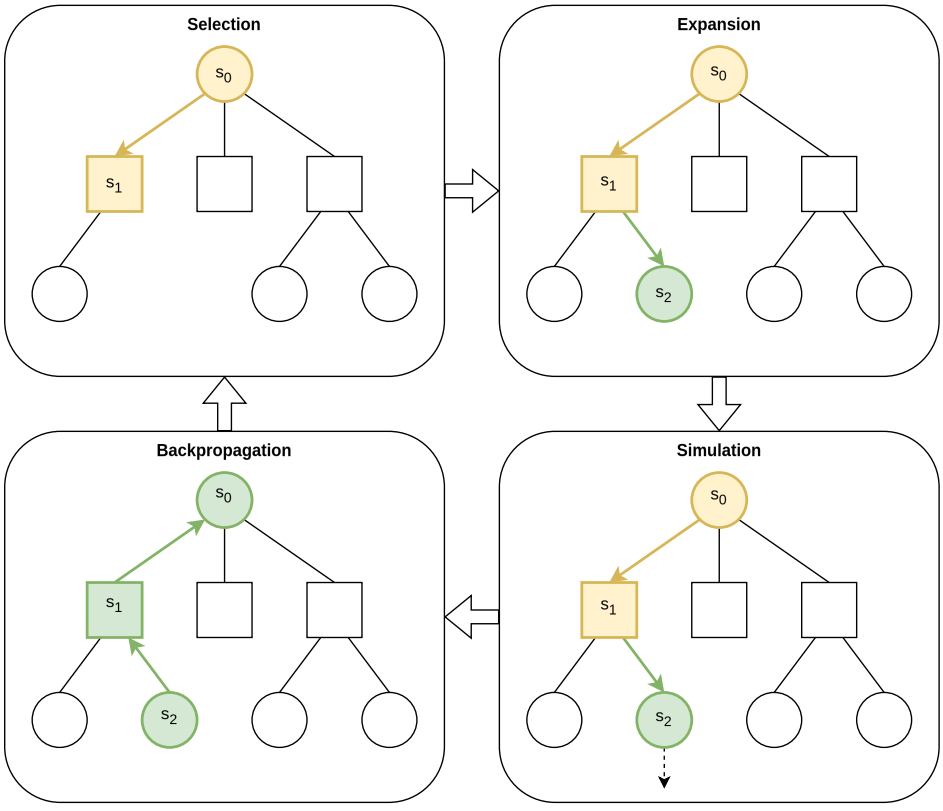


Figure 1: Individual steps of a single MCTS iteration. These are repeated in sequence until a given budget (time, number of iterations or other) is exhausted.

UCT formula.

$$\mathrm{UCT}(p,n) = Q(n) \times c\sqrt{\frac{\ln V(p)}{V(n)}} \tag{2}$$

with p the parent of n, Q the estimated reward and V the number of visits of a node. c is a parameter controlling the exploitation-exploration balance.

Another approach incorporates All Moves As First (AMAF) heuristics (Helmbold & Parker-Wood, 2009) to select nodes. Assuming that moves commute during search—a move that is good in one state is good in many others—it makes sense to keep track of the expected reward signal received after playing each move, regardless of the state they originated from. AMAF thus keeps one or multiple tables associating actions with their expected reward, using said tables to inform the selection process. Selection strategies incorporating AMAF values are often referred to as Rapid Action Value Estimation (RAVE) algorithms (Gelly & Silver, 2011), of which multiple variants have been devised depending on the number of tables stored and the way they are accessed (Sironi & Winands, 2016) (Figure 2). 3 presents the formula used in the most general variant, GRAVE (Cazenave, n.d.), to score individual children of a node, with the child maximizing $\mathrm{GRAVE}(n)$ being selected.

Generalized Rapid Action Value Estimation (GRAVE) formula.

$$\begin{aligned} \mathrm{GRAVE}(n) &= (1-\beta_{n,m})Q(n) + \beta_{n,m}\mathrm{AMAF}_{\mathrm{ref}}(m) \\ \beta_{n,m} &= \frac{\mathrm{AMAF}_{\mathrm{ref}}(m)}{\mathrm{AMAF}_{\mathrm{ref}}(m) + V(n) + (\mathrm{bias} \times \mathrm{AMAF}_{\mathrm{ref}}(m) \times V(n))} \end{aligned} \tag{3}$$

with n the node being scored, $Q(n)$ the current reward estimate, $V(n)$ is the number of samples and $m \in A$ the action played to reach this node. bias controls the reliance on AMAF values, while $\mathrm{AMAF}_{\mathrm{ref}}$ designates the AMAF table stored in the closest ancestor node with more than ref visits.

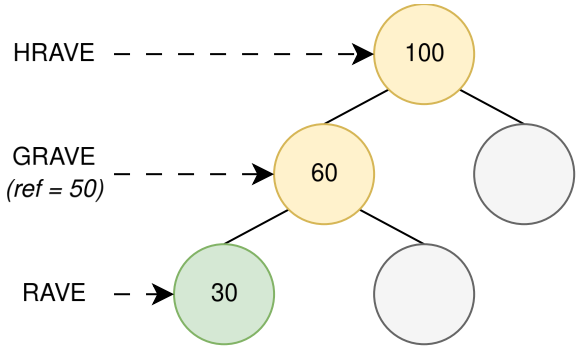


Figure 2: Nodes from which AMAF statistics are referenced during selection in the green node for the different RAVE variants (Sironi & Winands, 2016). Previously selected nodes are highlighted in yellow, and visit counts are indicated in the nodes. ***RAVE*** *only stores AMAF heuristics within leaf nodes;* ***HRAVE*** *stores a single table of AMAF heuristics for the entire tree; and* ***GRAVE*** *stores a table of AMAF heuristics in each node. GRAVE is a generalization of HRAVE and RAVE: using* $\mathrm{ref} = 1$ *emulates the behavior of RAVE, while* $\mathrm{ref} = \infty$ *provides the same behavior as HRAVE.*

The default playout strategy used in MCTS algorithm is to sample random continuations of the game: starting from the newly expanded node, random actions are played until a terminal state is reached. This has the advantage of not requiring any additional heuristic, but leads to noisy and unreliable backpropagated values due to the large number of possible continuations in most states—most of which are unlikely to be relevant. Methods to direct playouts exist, with the example of Move Average Sampling Technique (MAST) (Finnsson & Björnsson, 2010a) which combines reward estimates on moves and an ε-greedy selection strategy within playouts, or the use of weighted patterns (Gelly & Teytaud, 2006). State-of-the-art agents rely on neural networks capable of producing reliable signals taking into account long-term rewards (Silver, Schrittwieser, et al., 2017).

2.3. Heuristics for game-playing

Search algorithms by themselves often prove insufficient in games with large state spaces. In some cases, such as depth-first algorithms, they are required in order to avoid visiting the entire state space—an intractable endeavour in many board games.

To circumvent these issues, search algorithms employ heuristic functions. Such functions appear in the form of (1) state evaluation functions, which estimate reward signals in non-terminal states; and (2) policy functions, which estimate reward signals on actions rather than states. Historically, such heuristics were written by hand, encoding a set of rules to estimate the long-term value of states from positional features. These often required the intervention of experts to guide programmers on which features to look for, or how their relative value should be encoded.

Modern game-playing agents instead train deep neural networks to estimate long-term rewards and policies through reinforcement learning (Sutton & Barto, 2018) and expert iteration (Anthony et al., 2017). Agents trained through self-play to learn the expected reward from a given game state, as well as approximate a policy from MCTS, were shown to outperform human players in complex games where handcrafted heuristics are difficult to design (Silver, Schrittwieser, et al., 2017).

3

A cognitive view of human play

The psychology of experts in the game of Chess has been a subject of study starting from the late 19th century, with the pioneering work of Alfred Binet in "Psychologie des Grands Calculateurs et Joueurs d'Échecs" (Binet, 1894), treating on the ability of expert chess players to remember complex positions with little exposure. This work was later extended and refined by the foundational thesis of Adriaan D. de Groot, "Het Denken van den Schaker" (often translated as "Thought and Choice in Chess") (De Groot, 1946). This work has introduced many key ideas related to memory and planning in Chess experts. While many advances in the study of human game playing have been made since, the work of De Groot is still often cited as the canonical reference when mentioning the psychology of players. Game playing is also related to more general fields of study in cognitive science, with the most prominent being that of decision making.

Beyond their strategic and competitive aspects, multiplayer games are—perhaps mainly—social activities. This social context plays a key role, yet is often overlooked when discussing games through cognitive or programmatic lenses. Players cheat, lose on purpose, mock their opponents or are merciful to them depending on a variety of factors external to the game itself.

This chapter summarizes the results of empirical studies on the mechanisms related to how humans approach games, how they perceive, interact and learn with them. Sources on the matter can be broadly categorized as such: (1) cultural studies into the social nature of play; (2) psychological research on high-level features of decision-making and strategic planning in board games; and (3) supporting evidence in the form of cognitive approaches analyzing the biological processes involved.

We synthesize these elements into four interrelated dimensions: how players internally represent and memorize moves and game states (Section 3.1), how they approach the "Choice of Move" problem (Section 3.2), how knowledge generalizes across games (Section 3.3), and finally the broader motivations and social contexts shaping how and why humans engage with games (Section 3.4).

We acknowledge that a large part of this chapter's content is sourced from the book "Moves in Mind: the Psychology of Board Games" (Bart, 2012), a comprehensive resource on the matter for readers interested in delving deeper into this subject. The Ludus Ex Machina project's study of the social aspects of games in ancient Rome was also foundational in the writing of the related section.

3.1. Memory and representation

The ability of expert Chess players to play blindfolded was initially theorized a feat of visual memory (Binet, 1894). This was later challenged by showing that players relied on structured and efficient internal representation of game states as meaningful "chunks" of spatial relations rather than precise piece-square mappings (Chase & Simon, 1973; Gobet & Simon, 1998). This was done by measuring the ability of Chess players to reconstruct positions shown for a few seconds from memory. Some of these positions were sampled from real games, while others were crafted by randomly placing pieces. One notable result was the fact that expert players, while more capable of recalling complex Chess positions shown for mere seconds, did not perform better than less skilled players at reconstructing random assortments of pieces. This rules out the idea that visual memory plays a key role in the recall capacity of expert Chess players, and suggests that they instead develop an efficient internal representation of game states.

Similar results have later been generalized with the same experimental protocol to other board games such as Go (Reitman, 1976) or Othello (Wolff et al., 1984). However, it was noted that players had more difficulty reconstructing shown positions than in Chess. This might be explained by the higher volatility or complexity of game states. The most striking counterexamples are Bao and Awele, in which these results were never replicated (de Voogt, 1995). Outside of Chess, players were instead observed to rely on pattern recognition and structural qualities of the position.

While we found no study assessing the number of different board state representations players are able to retain in working memory at once, one can turn to simultaneous exhibition matches to estimate this number. In such exhibitions, one player is matched against multiple opponents, and is required to play against all of them simultaneously. Accounting for balanced skill between players, Garry Kasparov won against five grandmaster players simultaneously in 1988. Other simultaneous games with similar conditions have been played, hinting that handling five boards simultaneously could be a realistic estimate for the working memory capacity of expert Chess players. Similar results have been obtained more generally, with the "magical number seven" often being cited as an estimation for the number of representations humans are able to recall in working memory (Miller, 1956). Similar critiques to those targeted at said study also apply here: this does not account for any manipulations of the representation, or the recall time allowed before playing a move against each opponent in simultaneous exhibitions. Following results in Go, Othello or Bao for recall capacity, we can assume that this number is lower in more complex or volatile games.

The question of how players observe board states to construct this internal representation was studied through the recording of eye movements when presented a position. Experiments suggest that players' eyes follow "lines" along rule-related relations between pieces, rather than scanning the board as a whole. This forms a natural graph of interconnected pieces (Charness et al., 2001; Reingold & Charness, 2005).

3.2. Planning and decision-making

Players are known to perform some type of search in order to choose their next move. This "Choice of Move" problem sits at the intersection of two fields of study in cognitive science: decision making and forward problem solving.

Empirical studies on Chess players of various skill ratings show that they do not perform deep calculations, and that the depth of search does not significantly vary between strong and expert players, with both reporting that they calculate sequences of 5 or 6 moves on average (De Groot, 1946). The same study observed that the breadth of human search is also thin, with players considering only 3 to 4 moves—even though Chess has an estimated average of 31 to 34 legal moves per position. These findings strongly suggest that Chess players rely on an efficient pre-selection of moves, with stronger players capable of identifying the best moves much more accurately than novices. Search is only used to confirm or refute this intuitive selection.

In games with smaller branching factors (e.g. Bao or Awele), players rely more heavily on search (de Voogt, 1995), indicating that the selective nature of human players is dependent on the game's complexity.

The theory that search plays a secondary role to intuition is also backed by Magnetic Resonance Imaging (MRI) experiments in the games of Chess (Atherton et al., 2003) and Go (Chen et al., 2003), which show that areas of the brain typically linked to logical inference are less active than those related to spatial reasoning when faced with "Choice of Move" problems.

These empirical findings resonate with multiple theories and frameworks of general decision making. The interaction between two systems, one based on intuitive evaluation of options through learned patterns (System 1) guiding a more logical inference based on a system of rules (System 2), echoes with the Double Process Theory of Cognition (DPTC) (Kahneman, 2011). The idea that players perform only a shallow search to cement their intuition also relates to the framework of bounded rationality (Schwarz et al., 2022), which challenges the idea that humans can be modelled as purely rational agents in decision-making tasks. Individuals are limited by the effort required to obtain information and evidence for their choices. Thus, they tend to search for a satisfying solution—according to multiple relevant factors—rather than an objectively "best" option. This behavior has been dubbed as "satisficing", and should be understood as a dynamic framework, heavily dependent on player motivation, stress, time constraints and perceived advantage.

3.3. Transfer of knowledge

Human players rarely specialize on playing a single game over the course of their lives, with the exception of experts and professional players. Yet, individuals who play games regularly tend to perform better than occasional players when picking up a new game. The key observation is that individuals gather knowledge in a way that is transferable to other domains. The foundational framework studying the question of how and why knowledge transfers is that of the **near and far transfer framework** (Mestre, 2000; Thorndike & Woodworth, 1901a; 1901b; Woodworth & Thorndike, 1901). The main observation stemming from this framework and related experimental data indicates that skills transfer to the extent that two domains share overlapping cognitive features. Near transfer refers to transfer between closely related domains, while far transfer describes cases of transfer between loosely related domains.

In the context of games, studies on expert players suggest that while near transfer often occurs, far transfer is rare (Sala & Gobet, 2017). One explanation is that expert performance relies heavily upon specialized representations and learned patterns that are not applicable to other games. However, it can be argued that more general ideas related to game playing—threat evaluation, planning, identification of zugzwang situations—may transfer across games. While near transfer is prevalent in games, far transfer likely only encompasses general board game knowledge. This paints a hierarchical view: general concepts (far transfer), game-style specific knowledge (near transfer), and finally game-specific knowledge.

3.4. Motivations and social context of play

While AI research on game playing caters to the idea of improving playing strength above all else, humans do not play solely to win. Game playing is an inherently social activity, at both a local (friends, family) and global (federations, online platforms) level. Decisions are made not only by considering rule-defined goals, but also account for relationships, identity and social norms (Crist et al., 2016; Crist, 2019; de Voogt, 2005; Stenros & Montola, 2024).

Games are also shaped by their material nature (Sicart, 2022). This enables players to bypass the rules and manipulate boards, dice, pieces and cards, enabling interactions difficult to replicate abstractly like cheating by loading dice, hiding cards or moving pieces while the opponent is not looking.

Ongoing interdisciplinary work in human-like play studying Ludus, a Roman Backgammon-like game, highlights some of these socially-shaped behavior through the lenses of historical and modern evidence (Figure 3). For example, players were known to cheat by misreporting dice outcomes or manipulating the odds by loading dice intentionally. Adapting their own perceived playing strength to match that of the opponent is also observed throughout periods, whether to entertain or avoid angering said opponent.

*Figure 3: Exemple evidence of the impact of material aspects of play in historical context. (**Left**) fresco from a Pompeii tavern wall showing an accusation of cheating ("I win", "It was a three not a two") (CIL IV, 03494 from the Caupona of Salvius (Pompeii VI, 14, 35-36); (**Right**) Amethyst intaglio set in a gold ring depicting spectators interacting with players. Paris, Département des Monnaies, médailles et antiques de la Bibliothèque nationale de France, Luynes 115. From Dasen 2020 p. 178, fig. 5 (Dasen, 2020).*

Collaborative play, where multiple players work together as one, is also common ground. In the context of chance games or gambling, numerous fallacies and psychological effects have been reported, and their psychological impact on gameplay habits was directly studied (Tversky et al., 1982).

3.5. Discussion

While reviewing research on the cognitive and social nature of play, we identified some important gaps in knowledge that warrant further study. Assessing the working memory capacity of players during and outside of search is of importance to provide evidence to compare the memory usage of computer programs to that of humans. We identify the complexity of states themselves, the expertise of the player, and the similarity between said states—whether they are separated by one or multiple moves—as parameters to take into account for such a study.

Empirical evidence shows that far transfer is overstated between games and other domains, but studies of near and far transfer between games themselves are still lacking. This aspect could be quantified by cross-referencing the evolution of individuals across multiple games. This cross-referencing is more difficult than for single games, since well-known players often specialize in one game while most individuals remain anonymous on online platforms.

In a similar way, the generalizability of the claims related to memory and planning in games remains questionable outside of a subset of widely studied games (Chess, Shogi, Go or Mancala games), and focuses on experts. Broader studies need to be conducted to assess whether these empirical findings remain true across games and skill levels.

Regarding these last two points, we plan to leverage the popularity and wide catalogue of games included in the Ludii Game Playing System (Piette et al., 2020) to create an online platform for General Game Playing (GGP) (Genesereth et al., 2005), allowing the collection of human-generated data across multiple games for a community of individuals. We hope to study aspects related to generalizability by assessing the parallel evolution of individuals across many games, and allow an easier setup for experiments related to players' approach to game playing as a whole, rather than within specific games.

4

Classification of human-like models

While human-like agents are an active area of research, there doesn't seem to be a clear consensus on how such agents should be defined and categorized. Two contrasting views are prevalent: a **functionalist view** (Section 4.1) of human-likeness, and one based on the concept of **cognitive plausibility** (Section 4.2). Placing these views under the umbrella term of "human-likeness" without further classification creates confusion regarding the scope of models and prevents informed comparisons between them.

These views also tend to reduce game-playing as a set of isolated decisions and ignore individuality: *given a state of the game, what move does the average player select*? This view of game-playing is blind to many factors contributing to a player's decisions, identified in Chapter 3. We specifically focus on two additional axes: **modelling range** (Section 4.3), the variety of human motivations, preferences and playing styles captured by a model; **generalizability** (Section 4.4).

Section 4.5 synthesizes these concerns in a comprehensive classification. This classification will be referred to throughout the rest of this thesis in order to scope and properly identify the applicability of the methods presented.

4.1. Behavioral alignment

Historically, assessing whether human-like qualities emerged in a computer program was achieved through the Turing Test (Turing, 1950). In its original formulation, the test involves three participants: a human judge and two observees, one human and one program. The judge interacts with the observees through a computer interface, with the goal of identifying the computer agent. This method of measuring the human-like qualities of a computer program leans heavily on a functionalist view of human intelligence by solely accounting for the differences in observed behaviors: a computer program is human-like if its behavior cannot be distinguished from that of humans. We refer to this broad category as **behavioral alignment**.

Many research efforts towards human-like game-playing agents have adopted a similar methodology due to its practicality in providing quantitative measures of behavioral

alignment (Chu-Husan et al., n.d.; McIlroy-Young et al., 2020; Turnwald & Wollherr, 2019). In most of these results, the goal is to measure how accurately an agent is able to predict or imitate the moves played by humans over many game states. We identify this quality as **single move alignment**.

Human players have been shown to manipulate sequences, accounting the opponent's moves and past decisions into account. Aligning an agent's behavior to such **Sequences of Moves (SoM) alignment** is a worthwhile pursuit, notably if psychological features like opponent modelling (Thagard, 1992) are taken into account.

Finally, some models do not focus on moves, but rather on the general qualities of the games played. For example, game designers who want to assess the average length or balance of their product might benefit from models that play in a way that matches these metrics with those derived from human gameplay. We identify this as **statistical alignment**.

Despite its practicality and versatility, using behavioral alignment as a definition of human-likeness by itself presents important limitations. It measures the capacity of an agent to imitate, rather than understand and model human behavior, as pointed out by the Chinese Room argument (Searle, 1986) in regards to the Turing Test. As previously discussed, human play is shaped by cognitive functions and limitations such as pattern-recognition or bounded rationality, yet behaviorally aligned agents can be designed without taking these traits into account (McIlroy-Young et al., 2020). Behavioral alignment often assumes the existence of a canonical human playing style, ignoring the wide range of playing styles spanning various skill levels, motivations or preferences. Methods that measure alignment to averaged human behaviors take the risk of producing synthetic play patterns which do not correspond to any coherent human playing style, raising questions as to whether they can truly be considered human-like.

4.2. Cognitive plausibility

In order to circumvent the limitations of behavioral alignment, we propose to complement it with the notion of **cognitive plausibility**. A model is cognitively plausible if the computational methods it uses could realistically be compared to cognitive functions of biological brains. This should be understood as a spectrum rather than a binary classification, and is evaluated subjectively in the absence of a comprehensive model of cognition. Appraisals of cognitive plausibility need to be discussed extensively and motivated through empirical results when available. For example, it can be argued that a search algorithm visiting only a dozen of game states is more cognitively plausible than one visiting millions of positions due to known representational limitations of our brains and phenomena of informational overload.

The work of Allen Newell in *Unified Theories of Cognition* (Newell, 1994) was one of the first to introduce criteria to properly define cognitive systems, while later work by Duch et al. (Duch et al., 2008) refined these criteria into a taxonomy based around two dimensions: memory and learning (Figure 4). In the latter, memory refers to learned intuitions rather than how it is understood from the standpoint of game-playing programs. Learning and memory mechanisms are intertwined with the former modifying the latter, and the latter biasing the former.

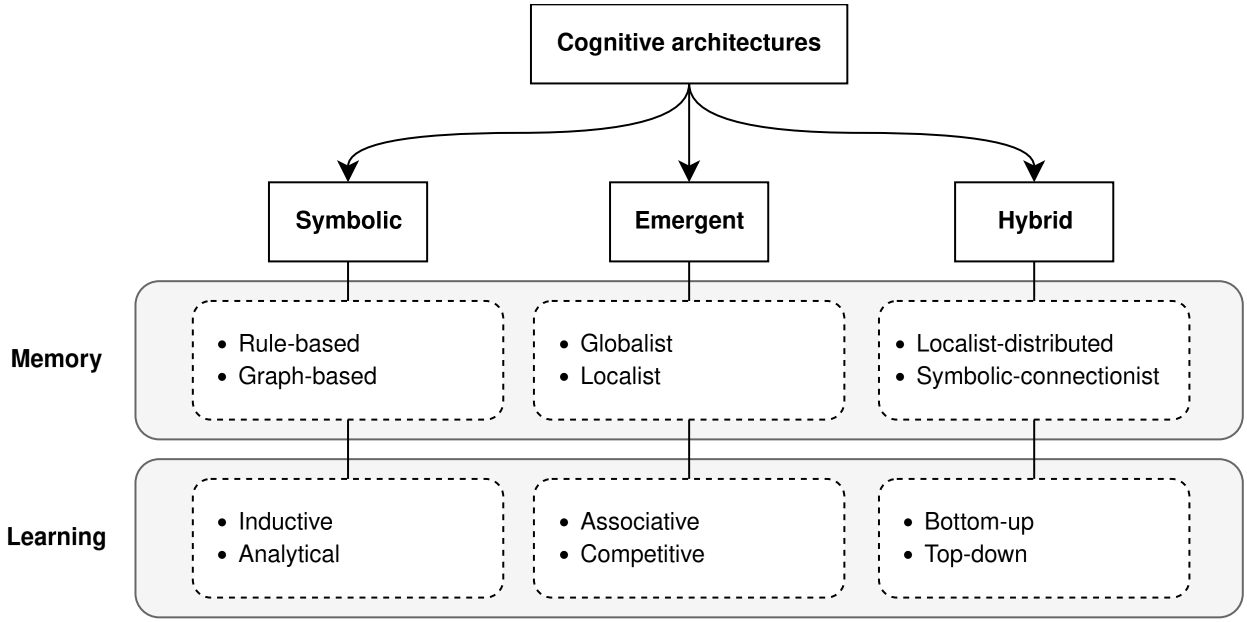


Figure 4: Simplified taxonomy of cognitively plausible computerized models (from (Duch et al., 2008)).

One issue with this presented taxonomy is that it does not take into account the biological feasibility of the algorithms used. For example, associative learning is usually implemented using backpropagation (Goodfellow et al., 2016), which has been criticized as being possibly ill-fitted to our current knowledge of the brain's learning mechanisms (Lillicrap et al., 2020). We argue that coherence with the biomechanics of the human brain needs to be a core part of cognitively plausible systems.

While cognitive plausibility and behavioral alignment might seem tightly linked, it is important to treat them separately. Some definitions of human-likeness involve cognitive plausibility as a prerequisite (e.g. biological naturalism (Searle, 1986)), but there is no conclusive evidence showing that behavioral alignment can only arise from cognitively plausible systems or that cognitive plausibility is a sufficient condition for behavioral alignment. Systems based on imitation learning, for example, are able to replicate human behavior without modelling or understanding the cognitive systems that it originates from.

4.3. Modelling Range

Along with cognitive plausibility as a means to alleviate the blind spots of behavioral alignment, modelling range addresses the variety of playing styles and goals. While high-level features of cognition might generalize across players—although accounting for pathologies and/or neurotypes might weaken this assumption—it is clear that there is no unique "human playing style" from which to derive a canonical behavior.

We propose that human-like models should take into account the **granularity** against which they measure behavioral alignment. Most of the models discussed earlier approach the problem over an entire population of players by comparing a model's behavior to that of an averaged human policy. This policy is obtained from a population regardless of playing style or individual motivations. This might lead to synthetic play patterns not corresponding to any individual, but rather to a hypothetical crowd voting to select the next move to play. A first refinement would be to consider a singular playing style by grouping individuals based on the similarity of their observed behaviors (e.g. opening preferences, skill rating, etc). Again, such models do not need to match the behavior of any single

individual, but should output a playing style coherent with the clustered players. Finally, models attempting to mimic a single player should be observed to have a coherent playstyle similar to said player.

For playstyle-wide and single-player granularities, we also incorporate a notion of versatility. Human-like models that can be tuned in order to match a variety of playing styles/players should be differentiated from those specifically created to mimic unique playing styles/players.

Granularity can be seen as various levels of clustering on whichever dataset is compared against to measure behavioral similarity. Population-wide models use only one cluster (the entire dataset), while single-player models use individual data points, with playstyle-wide models sitting somewhere in between.

4.4. Generalizability

A final axis along which to classify human-like game-playing systems is their ability to generalize knowledge across multiple games. This relationship is referred to as the **transferability** of knowledge from one domain A to another domain B, which we denote $A \underset{k}{\rightarrow} B$. Intuitively, all pairs of domains are not equal in this regard in the context of the near-far transfer framework (Mestre, 2000): more knowledge might be transferred between Chess and Shogi, which share similar mechanics in the form of goals and moves, than between Chess and Go. Conversely, this relationship is not necessarily commutative, and it is possible that knowledge transfers better in one way than the other i.e. $A \underset{k}{\rightarrow} B = B \underset{k}{\rightarrow} A$ is not necessarily true.

For human-like game-playing agents, this transferability maps directly to the number of games they are able to play while maintaining their human-like qualities, without requiring drastic changes to their implementation. This definition is admittedly loose, however it catches some cases like neural networks that require changes to their input space in order to adapt to new games.

In the best case, we could hope for models that do not require extensive retraining to learn how to play **many games** (Rautureau & Piette, 2025). This point is especially relevant to the field of GGP (Genesereth et al., 2005), which aims to create agents able to play many games by only providing them with their rules, thus not enabling the use of domain-specific knowledge. While some time is given to these systems to familiarize themselves with each game, a program requiring long training steps for each individual game would be impractical.

While distinguishing human-like agents able to play many games from those that can only play one is necessary, agents able to play a class of related games are also of interest. If agents able to play many games are perhaps a long-term goal, agents able to play a **class of similar games** (e.g. war games like Shogi and Chess, or connection games like Connect-4 or Renju) are likely to be a first stepping stone.

Most game-playing agents are fit to play a single game, especially when using heuristics which often require domain-specific knowledge. The design of general heuristics has been studied in the context of game-playing agents (Soemers et al., 2023; Stephenson et al., 2021), and is a promising direction to extend methods for human-like agents to the domain of GGP (Genesereth et al., 2005; Rautureau & Piette, 2025).

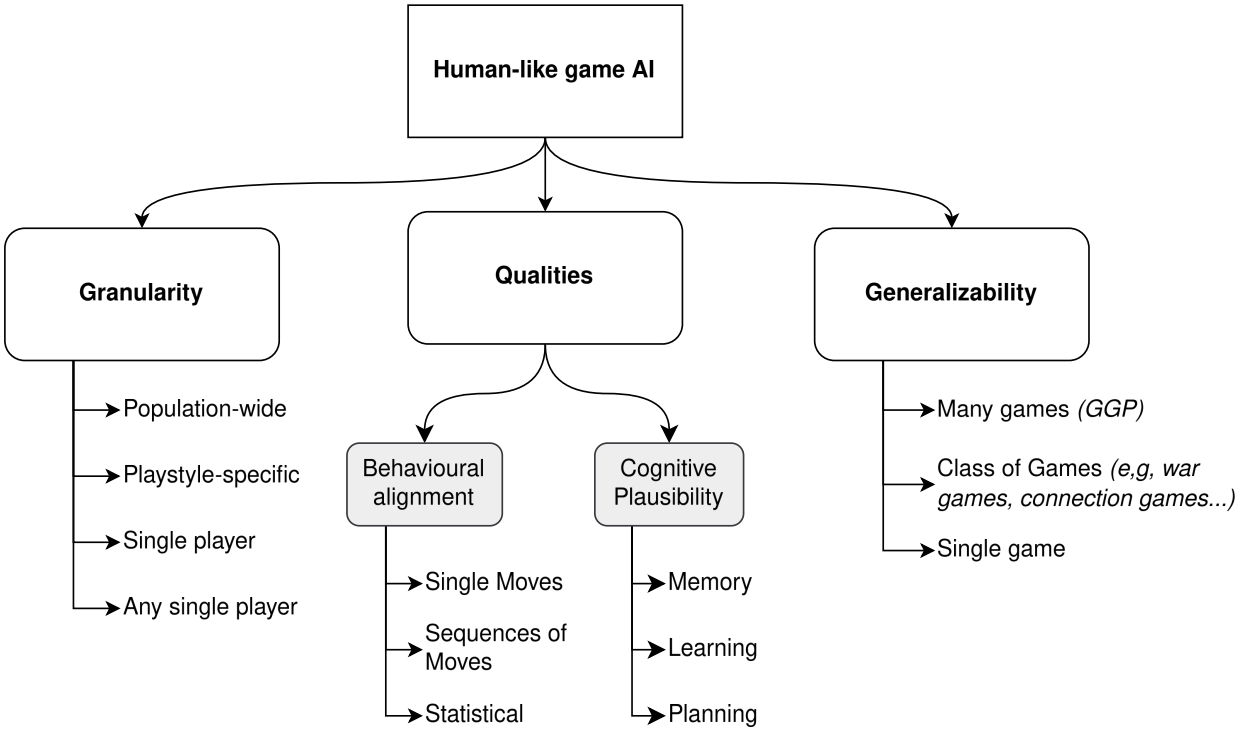


*Figure 5: Classification of human-like agents across three axes: **granularity** for the breadth of behaviors they capture, **qualities** for behavioral alignment and/or cognitive plausibility, and **generalizability** for their ability to transfer knowledge across games.*

4.5. A classification of human-like models

We synthesize the qualities outlined in this chapter as a general classification of human-like models based on granularity, qualities, and generalizability (Figure 5).

This classification enables a more precise discussion around human-like models, and identifies some areas for future research in the field, like generalizability, Sequence of Moves alignment and model granularity. For example, the current state-of-the-art models in Chess based on imitation learning (McIlroy-Young et al., 2020; Tang et al., 2024) fit the category of single-game, population-wide, single-move alignment models.

This thesis mainly focuses on cognitively plausible methods applicable to many games. The main axis of this work is to bring the algorithmic core of game-playing agents closer to human cognitive functions and limitations. Nevertheless, we also touch upon measuring behavioral alignment in various forms in Chapter 5. We plan to work towards a characterization of the relation between cognitive plausibility and behavioral alignment in future experiments to answer a key question: *do cognitively plausible models tend to be more behaviorally aligned in games*?

5

Measuring behavioral similarity

We now turn to methods to measure the behavioral alignment of game-playing agents given the functionalist framework highlighted in Chapter 4 outside of the Turing test (Turing, 1950).

The advent of large open datasets recording the games played between human players[1][2] enables the use of statistical methods measuring the divergence between an agent's behavior and that of human players. We first present the framework within which we ground these statistical methods (Section 5.1). Given our classification of granularity, we identify three different types of methods, depending on the number of artificial and human policies compared: **one-to-one** (Section 5.2), **one-to-many** (Section 5.3), and **many-to-many** (Section 5.4). Current approaches to statistical measurements for behavioral alignment in games cover one-to-one scenarios (McIlroy-Young et al., 2020; Tang et al., 2024), while research on the remaining two remains scarce.

This chapter aims to present metrics, discuss their applicability, caveats and advantages. Section 5.5 then recontextualizes results on the behavioral alignment of Chess programs. We discuss future work for a broader range of metrics in Section 5.6.

5.1. Framework of behavior

Within an MDP, an agent's behavior is modelled through its **policy** π. In practice, only observations of this behavior are available. Due to many games having large state spaces, it is unlikely for players to have visited all states at least once.

We formalize these observations as a multiset $O_\pi = (U, m)$, with U the underlying set and $m : U \to \mathbb{N}$ the multiplicity of each element. For deterministic games, we define $U = S \times A$. O_π can be thought as a dataset counting the number of times an agent was observed to take action a in state s.

[1] Lichess database: https://database.lichess.org
[2] GNU Shogi database: https://japanesechess.org/gsdb/

These observations can be used to derive an estimation $\hat{\pi}$ of the agent's true policy π (Equation 4). In practice, some filters must be applied to these observations to reduce noise, for example by ignoring states s with $\sum_{a \in A} m(s,a) < t$ observations.

$$\forall (s,a) \in S \times A, \hat{\pi}(s,a) = \frac{m(s,a)}{\sum_{a' \in A} m(s,a')} \tag{4}$$

For artificial agents, it is sometimes possible to extract an exact policy e.g. the output of a policy network if no further stochastic operations are applied to it. In practice, most game-playing agents only return a "best move" for a given position. In such cases, repeated queries should be performed to account for any stochastic processes at play—temperature applied to policies, ε-greedy strategies, random MCTS playouts, etc.

We can additionally define a simple measure of the decisional variance—how much the moves selected vary within states—of an observed policy O_π as the normalized informational entropy (Equation 5) of its derived policy $\sigma(\hat{\pi}) = \frac{1}{|S_\pi|} \sum_{s \in S_\pi} \frac{H(\hat{\pi}(s))}{\log_2 |A_s|}$, where S_π is the set of states in O_π.

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log_2 p(x) \\ &= \mathbb{E}(I(X)) \\ 0 \leq H(X) &\leq \log_2 |X| \end{aligned} \tag{5}$$

Decisional variance by itself can be used to differenciate policies based on how focused they are. Most artificial agents are deterministic if none of their parameters (searching time, maximum depth, etc) are changed, which corresponds to a decisional variance of 0. In contrast, we can expect human players to have a positive decisional variance.

5.1.1. Distances between policies

Measures of behavioral alignments rely on computing divergence between observed policies. Studies on human perception show that measures of similarity are a better surrogate than those of distance (Fechner, 1948) and that our perception of similarity and/or distance is not linear. For example, the Magnitude Effect (Kahneman & Tversky, 2013) states that humans tend to have diminished sensibility to differences in larger numbers, which is aligned with the Weber-Fechner law (Fechner, 1948) in psychophysics, stating that the relation between perception and stimuli is logarithmic.

In the case of policies, divergence functions are composed of two parts: a divergence on probability distributions between each state policy, and an aggregation of these divergences over all states. Distances between probability mass functions are a well-studied problem, with the main difference between them being their interpretation. A summary of some of these functions is given in Table 1.

5.2. One-to-one behavioral alignment

One-to-one measurements are suited for population-wide or player-specific behavioral alignment. The most common example in the literature is **move matching** (McIlroy-Young et al., 2020). It measures the capability of an artificial agent to predict actions taken by human players, and has been notably used to measure the human-likeness of the Maia family of Chess engines (McIlroy-Young et al., 2020; Tang et al., 2024). It can be formulated as a computation of the expected probability that an agent will select the same actions

5

Name and formula	Properties	Interpretation
Wasserstein metric (Kantorovich, 1939) $W_p(X,Y) = \inf_{\gamma \in \Gamma(X,Y)} \left(\mathbb{E}_{(x,y)\sim\gamma} d(x,y)^p\right)^{\frac{1}{p}}$	Metric	Effort to transform one distribution into the other
Jensen-Shannon distance (Lin, 1991) $\mathrm{JSD}(X,Y) = \sqrt{\frac{1}{2}(\mathrm{KL}(X \mid M) + \mathrm{KL}(Y \mid M))}$ with $\mathrm{KL}(X,Y)$ the Kullback-Leibler divergence and $M = \frac{1}{2}(X+Y)$	Metric, bounded by 1	Information theoretic (mutual information, expected surprisal)
Bhattacharyya distance (Bhattacharyya, 1946) $D_B(P,Q) = -\ln(\mathrm{BC}(P,Q))$ with $\mathrm{BC}(P,Q) = \int_{\mathcal{X}} \sqrt{p(x)q(x)}dx$	Not a metric, unbounded	Likelihood that two distributions will yield different values
Hellinger distance (Hellinger, 1909) $D_H = \sqrt{1 - \mathrm{BC}(P,Q)}$	Metric, bounded by 1	Same as Bhattacharyya

Table 1: Functions of distance between probability distributions.

following policies π and $O_{\pi'}$ (Equation 6). $O_{\pi'}$ is a set of population-wide observations. A notable flaw of move matching is that it does not measure the divergence between each state's probability distribution over actions. For example, the move matching score of a probability distribution against itself is not 100%. More generally, move matching favors "best move" policies, even when the target policy showcases multiple choices (Table 2).

$$\mathbb{M}(\pi, O_{\pi'}) = \frac{1}{|O_{\pi'}|} \sum_{(s,a)\in O_{\pi'}} m(s,a) \times \hat{\pi}(s,a) * \pi(s,a) \tag{6}$$

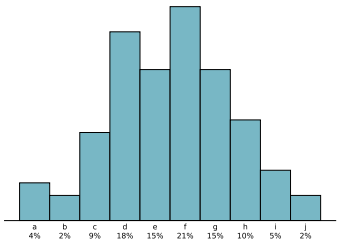
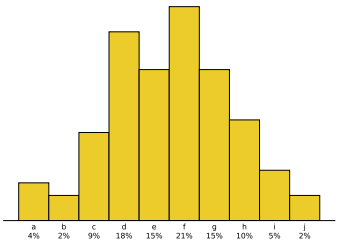
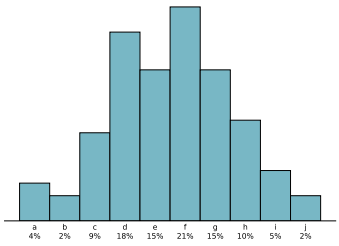
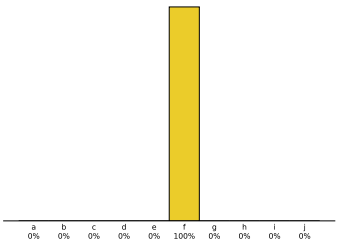
Observed policy $\hat{\pi}$	Agent policy π	$\mathbb{M}$ (*higher is better*)	D_B (*lower is better*)
		14.45%	0
		21%	0.78

Table 2: Minimal example showing that move matching ($\mathbb{M}$) will always favor "best move" policies, even when the observed target policy might consider multiple options. This is compared to using the Bhattacharyya distance D_B on state policies.

Playstyle Distance (PD) (Lin et al., 2021) is an improvement over move matching. While originally introduced to be applied to video games, it can be used as-is for tabletop games and general MDPs. PD is measured by isolating a subset of states common to two observed policies O_π and $O_{\pi'}$, which we denote $S_\cap = \left\{ s \mid \sum_{a \in A} m_{\pi(s,a)} \neq 0 \wedge \sum_{a \in A} m_{\pi'}(s,a) \neq 0 \right\}$. For each common state $s \in S_\cap$, we then compute $d(O_\pi, O_{\pi'} \mid s) = D\left(\hat{\pi}(s), \hat{\pi'}(s)\right)$, with D a measure of distance between the probability mass functions of $\hat{\pi}$ and $\hat{\pi'}$ in state s. In the original paper, the 2-Wasserstein distance was used, but it can be replaced by any measure of divergence between probability distributions. We select the Hellinger metric (Hellinger, 1909) over Bhattacharyya (Bhattacharyya, 1946) due to it being bounded, which makes it behave correctly under summation.

The overall distance between O_π and $O_{\pi'}$ is then computed following Equation 7 by taking into account the distribution of samples over $S_\cap$.

$$\begin{aligned} \mathrm{PD}(O_\pi, O_{\pi'}) &= \frac{1}{2}\left(\hat{d}(O_\pi \mid O_{\pi'}) + \hat{d}(O_{\pi'} \mid O_\pi)\right) \\ \hat{d}(O_x \mid O_y) &= \mathbb{E}_{(s,a) \in O_y, s \in S_\cap}\left[d(O_x, O_y \mid s)\right] \end{aligned} \tag{7}$$

As it stands, PD does not define a metric space, even if D is a metric. It is symmetric ($\mathrm{PD}(A,B) = \mathrm{PD}(B,A)$) and follows the axiom that the distance from any observed policy to itself is 0 ($\mathrm{PD}(A,A) = 0$), but does not satisfy the triangle inequality ($\mathrm{PD}(A,C) \leq \mathrm{PD}(A,B) + \mathrm{PD}(B,C)$) or positivity (if $A \neq B, \mathrm{PD}(A,B) > 0$), as shown in Table 3. A specific case where PD satisfies all metric axioms is when the set of observed states is shared between all individual policies. This can be achieved in two ways:

- Computing a subset of states common to all policies, and ignoring all other states. In practice, this reduces the usable dataset too heavily in games with large state spaces. Taking into account symmetries and rotations might circumvent this issue.
- Assuming that missing states are associated with a uniform policy as a form of neutral element. This applies a lesser penalty when compared to state policies with high variance, but a higher one when compared to single-move policies.

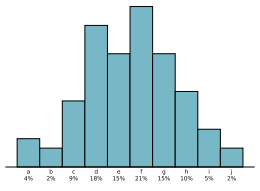
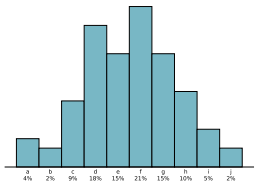
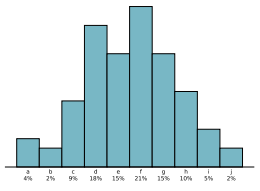
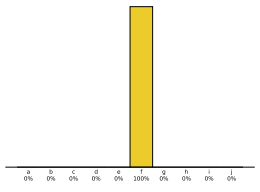
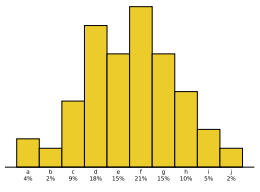
Agent	A	B	C
x			
y		None	

Table 3: Synthetic example showing that PD does not uphold positivity between different policies, nor the triangle inequality. Here, $A \neq B$ yet $d(A,B) = 0$, and $d(A,C) \geq d(A,B) + d(B,C)$. Note that this is due to B not defining a distribution over actions in state y, otherwise both properties would hold.

From a perceptual standpoint, applying a penalty when two policies do not visit the same states is a balancing problem. Players visit different states either due to differences in their preferences (e.g. in opening sequences), or due to factors external to their own decision process (e.g. the opponent's decision, randomness in stochastic games).

Playstyle Similarity (PS) (Lin et al., 2024) aims to improve PD by better aligning it with human perception: an exponential kernel $P(d) = \frac{1}{e^d}$ is used to emulate non-linearity in perception (Fechner, 1948), and the Bhattacharyya distance (Bhattacharyya, 1946) replaces the less perception-aligned 2-Wasserstein distance (Equation 8).

$$D_{\text{PS}}(A, B) = \frac{\sum_{s \in S_\cap} P(D_B(A, B))}{|S_\cap|} \tag{8}$$

In order to solve the issues that arise from using the intersection of states to measure distance, PS incorporates the Jaccard index (Equation 9) (Murphy, 1996). Integrating the proportion of common states over the number of total states observed in the policies is a sensible middle ground, only feasible because it defines a similarity rather than a distance.

$$J(A, B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} \tag{9}$$

Synthesizing these considerations provides the Playstyle Similarity metric Equation 10.

$$\begin{aligned} \text{PS}(A, B) &= J(A, B) \times D_{\text{PS}(A,B)} \\ &= \frac{\sum_{s \in S_\cap} P(D_B(A, B))}{|S_A \cup S_B|} \end{aligned} \tag{10}$$

5.3. One-to-many behavioral alignment

One-to-many behavioral alignment measures take into account individual policies rather than aggregating them as done in one-to-one measures. No such measure has been designed to our knowledge. We introduce the **Policy Outlier Detection (POD)** method, a many-to-one statistical test using distance-based outlier detection methods (Wang et al., 2019) to assess the likelihood of an observed policy O_π being part of the target population $\{O_0, O_1, ..., O_n\}$.

We use the Local Outlier Factor (LOF) algorithm (Breunig et al., 2000) in order to detect whether a data point is anomalous when added to the set. The novel point is marked as an outlier if its local density is lower than that of its k-nearest neighbors.

Formally, LOF defines a distance $d_k(A)$ from A to its k^{th} nearest neighbor. The set of k-nearest neighbors is denoted $N_k(A)$, and may contain more than k points in the case of ties between them. A reachability distance is then defined as $r_k(A, B) = \max\{d_k(B), d(A, B)\}$, and used to derive the local reachability density of $\text{lrd}_k(A) = \frac{|N_k(A)|}{\sum_{B \in N_k(A)} r_k(A,B)}$. Finally, we define the Local Outlier Factor (Equation 11) as the ratio between the average local reachability density of the k-nearest neighbors of A and its own local reachability density.

$$\text{LOF}_k(A) = \frac{1}{|N_k(A)|} \sum_{B \in N_k(A)} \left(\frac{\text{lrd}_k(B)}{\text{lrd}_k(A)} \right) \tag{11}$$

If $\text{LOF}_k(A) \leq 1 + \varepsilon$, A has a similar or higher density compared to its neighbors, and is thus not an outlier. If $\text{LOF}_k(A) > 1 + \varepsilon$, A is classified as an outlier. The value $\varepsilon \in \mathbb{R}^+$

controls the sharpness of the decision boundary. and should be fitted depending on the dataset used. The choices of both k and ε have to fit to the target dataset so that none of the target data points are considered outliers.

Any one-to-one distance metric can be used between policies, as LOF remains well behaved for distances that do not define a metric space as long a few violations of the triangle inequality arise.

This method has the advantage of providing a clear way to compare different agents: agents with a lower LOF will be considered "less of an outlier" than their peers. It also defines a decision boundary for human-likeness, dependent on the choice of parameters ε and k. Overall, it provides a general method for measuring behavioral alignment while accounting for individual behaviors rather than aggregated ones. Many other methods for distance-based outlier detection exist, and might be applied to the problem of one-to-many behavioral alignment tests.

5.4. Many-to-many behavioral alignment

Many-to-many tests are the only ones suited to measuring behavioral alignment for models attempting to imitate multiple playing styles or players. Their goal is to measure the alignment of a population of artificial policies produced by a model's various parameters against a population of human players.

One way to approach many-to-many measurements is to build upon one-to-many methods and aggregate the results of individual policies within the artificial population. One can measure the average LOF over a population of novel points, along with its standard deviation, to decide whether the model can be expected to produce human-like behavior over a range of parameters.

Another possible approach is the use of dimensionality reduction methods such as Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes et al., 2020) to embed policies from both populations into a low-dimensional metric space. These embeddings can be used to measure the distance between the distribution of data points in the artificial and target populations. This can be performed using the n-Wasserstein distance (with n the number of dimensions), for example. One issue with such pipelines is the difficulty of assessing the "loss of information" inherent to projecting policies into a space with lower dimensionality. It can however provide insights into the distribution of behaviors produced by the model and their relationship to individual human behaviors. This can help identify the fact that models might favor some playing styles over others.

5.5. Experiments

We now turn to recontextualizing previous results obtained by means of the Move Matching metric using PD, PS and POD. Experiments will be conducted on the game of Chess. We use the Lichess database[1] of games played in the year 2017.

To focus on decision making rather than learned lines, we ignore the first 10 moves played in each game, approximately corresponding to the opening sequence. Time controls that allow less than 3 minutes per game were also removed. Finally, we define a dataset of observed policies for each player and each 100 rating point bin between 1200 and 2000. This gives us 9 distinct datasets in total, and avoids comparing players of different skill levels. Due to the evolving nature of ratings, a single player may appear in multiple datasets.

Only states with more than 3 observations were retained in an effort to balance noise and the amount of individual policies retained. Applying these filters results in datasets ranging from $9,743$ to $32,419$ individual player policies per rating bin ($\mu \approx 22,832$), and between $21,603$ and $151,178$ unique states ($\mu = 94,904$).

We compare Stockfish 18[3], the strongest Chess engine with a rating of 3642 ELO according to the CCRL website at the time of writing[4] ; and Maia Chess, a Chess engine trained through imitation learning on a similar dataset as the one used here. Stockfish uses a fixed depth of 9 to replicate the experimental setup used in the Maia Chess paper (McIlroy-Young et al., 2020). For distance measures between agent policies and human policies, it is assumed that the agent has visited the exact same states as the human player. Since the states visited within by the agent do not arise from real gameplay, they are not indicative of any playing style considerations.

We first compute the PD, PS and Move Matching of each agent against the average human policy derived from each dataset (Figure 6). The results obtained by both PD and PS are similar to those obtained using Move Matching. This can be explained by the fact that all compared agents have a decisional variance of 0, meaning that the Bhattacharyya and Hellinger distance are essentially equivalent to Move Matching. However, the decisional variance of the human policies (Figure 7) is relatively high, meaning we can hope to improve PD and PS by incorporating variations in an agent's playing style. Additionally, we observe a striking difference in decisional variance between the averaged policy and the averaged decisional variance of individuals. In practice, this indicates a behavioral difference between agents following these policies, as the averaged human policy has a less "focused" playing style compared to individuals. This motivates the idea that averaged policies might not retain the "human-like" qualities of their parts.

In order to motivate one-to-many measures as more accurate than one-to-one measures for population-wide modelling granularities, we also measure the average distance between each agent's policy and individual humans' policies. We observe that while global trends are retained (e.g. Stockfish performs better as rating progresses), the standard deviation of these distances is high no matter the method used, accounting for around 0.3 on average. This is a strong indication that the datasets present a large variance which is not captured by one-to-one measures against an averaged policy.

We then turned to measuring the LOF against the same dataset of individual policies. ε and k parameters were fitted to each rating bin by finding the minimum values such that $\mathbb{E}_{\hat{\pi} \in \hat{\Pi}}[\mathrm{LOF}_k(\hat{\pi}) < 1 + \varepsilon] \leq 0.05$ (only 5% of the dataset is considered outliers at most), and so that the variance of LOF scores within the dataset is below 0.25. We did not manage to obtain results for LOF on the dataset used. The issues encountered were due to the lack of a true metric property on any of the available one-to-one distances, leading to $\approx 58.67\%$ of individual policies being considered outliers at best ($k = 105$, $\varepsilon = 1$). Using PD along with a penalty for non-shared states between policies heavily favored artificial policies due to them not being affected by said penalty.

The experimental code for this section can be accessed on https://codeberg.org/thinkingrocks/policy_outlier_detection.

[3] https://stockfishchess.org/

[4] https://www.computerchess.org/

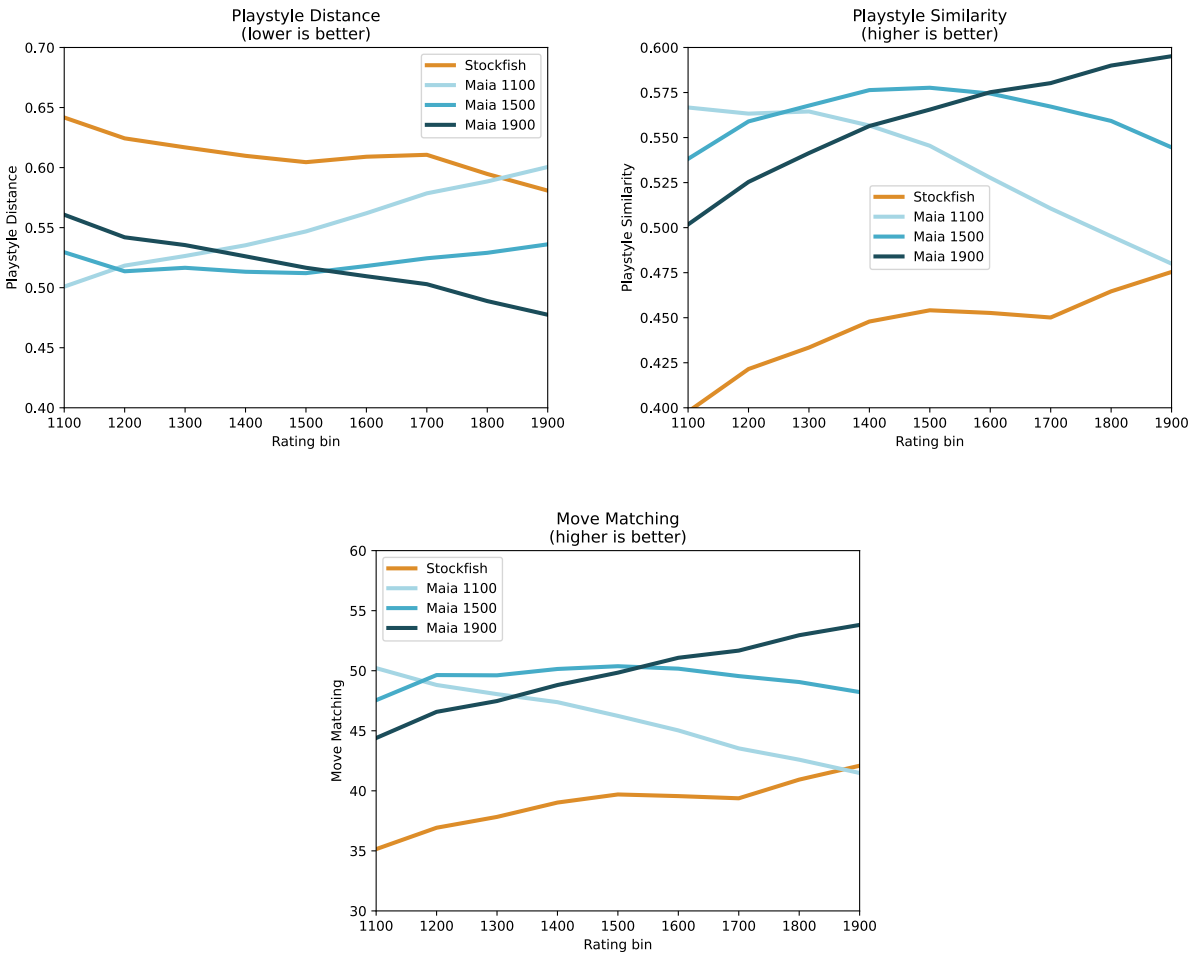


*Figure 6: PD (**left**), PS (**right**) and move matching (**bottom**) of compared agents to the averaged human policy across multiple rating bins.*

5.6. Discussion

This chapter provided an overview of 5 statistical tests for behavioral alignment, 4 of which were retained as being usable in practice. The main contribution of this chapter, POD, was not applicable due to the lack of a proper metric between policies. We thus identify the development of better-behaved one-to-one metrics as a focal point of future research to move past one-to-one measurements on averaged policies.

It is important to note that these statistical methods may not accurately reflect human perception of differences in playing style and behavior. Out of the methods discussed here, PS is the most cognitively motivated metric and should be preferred when possible. The experiments conducted also recontextualize previous results from the current state-of-the-art human-like model in Chess, Maia Chess. While they showcase great results as population-wide models, our experiments show that great improvements over this approach can be made by integrating variance into the move selection of programs. We plan to reevaluate the performance of imitation learning models by accounting for their full policy and stochasticity.

Methods for SoM alignment are scarce compared to single-move ones. Although not discussed here, they are paramount to capturing strategic coherence and interaction with the opponent in human-like agents. We plan to work on designing metrics that take into account trajectories within the MDP framework rather than individual transitions. Other

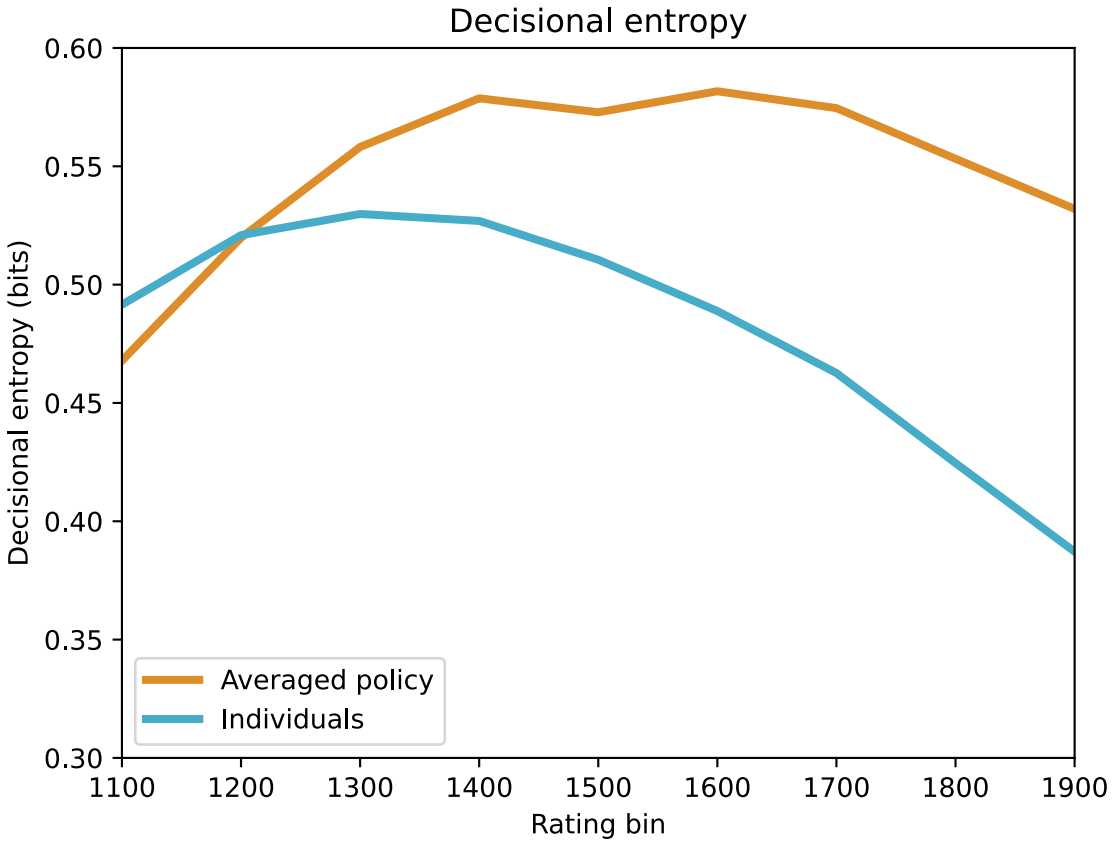


Figure 7: Decisional variance of players across rating bins. We show both the decisional variance of the averaged policy, and the average decisional variance of individual players.

approaches inspired by stylometry are noteworthy for SoM alignment (McIlroy-Young et al., 2021).

While human-like models in GGP are a long-term goal, we rapidly discuss an experimental setup for measuring generalizability. As a model acquires more knowledge in one domain, we can show that this knowledge transfers to other domains by measuring correlations in performance-related metrics. In the simplest case of playing strength, we can correlate improvements in any number of domains as a function of training time in a single domain. Non-zero correlation coefficients indicate that knowledge must transfer in some way. Note that negative correlation coefficients also count as a form of transfer, although it has a negative impact in playing strength for other games. The only case where transfer does not happen is in the case of a zeroed coefficient.

While this method is sufficient if only playing strength is taken into account, general human-like models should also remain behaviorally aligned across domains. This is especially important for confidently assessing that these models behave similarly to human players, even when little to no human-generated data exists (e.g. for ancient games).

6

Cognitively plausible search

Following discussions of definition and measures, we move towards exploring cognitively plausible search algorithms for game-playing programs as a first example of human-likeness beyond behavioral similarity. The methods presented in this chapter focus on approaching realistic cognitive constraints and mechanisms, rather than achieving behavioral alignment. The assessment of a relation between these two qualities is left as future work, as it would not be realistic to perform comprehensive experiments within the time scale of this thesis.

Cognitively plausible search algorithms are a less extensively researched area of human-like agents compared to heuristics. We observe striking differences between the way humans and machines search: while machines are able to traverse game graphs at rates of millions of states per second, human players require more mental effort to simulate sequences of moves in their mind. The ways in which brain and machine store information are also incomparable, with one using abstract representations of varying efficiency, while the other can be precisely measured in bits. In this chapter, we consider a node stored within the game tree as an atomic piece of information.

While the speed, depth and memory budget of search algorithms can be tuned in a number of ways, these methods have been observed to drastically reduce the coherence of a program's playing style (Nau, 1983). This highlights another difference between the way humans and computers search: players do not stop searching when hitting a memory, depth or time bound, but rather when they find a satisfying solution (Schwarz et al., 2022).

This chapter aims to bring search algorithms closer to human cognition regarding memory (Section 6.1), selectivity (Section 6.2) and bounded rationality (Section 6.3). This chapter is concluded by a set of experiments conducted on the game of Go (Section 6.4), and a discussion of these results along with potential future research directions (Section 6.5).

6.1. Memory constraints

We first present various modifications that can be made to any MCTS algorithm with the goal of reducing its memory budget while maintaining playing strength. This is done through cognitively plausible mechanisms, inspired by the Atkinson-Schiffrin model of

working memory (Atkinson & Shiffrin, 1968) (Section 6.1.2) and the intuition that humans tend to solve recursive subgoals rather when approaching forward problem solving (Hay et al., 2014) (Section 6.1.1). These considerations are brought together into a cohesive framework based on MCTS variants already used in game-playing agents. The overall algorithms are thus similar to current state-of-the-art search algorithms, but tweaked to gracefully handle limited memory capacity.

Throughout this chapter, we use the number of nodes N stored in the game tree as a substitute for cognitive load. Furthermore, we will only concern ourselves with asymptotic memory usage to avoid taking into account data structure-dependent memory footprints, while not ignoring asymptotically impactful data on which some algorithms depend (e.g. heuristics for each possible move in Rapid Action Value Estimation algorithms).

6.1.1. Two-level search

A first approach to reduce the memory requirements of MCTS algorithms is to use a two-level scheme (Breuker, 1998), partitioning the stored search tree into a permanent top-level tree, and a temporary second-level tree. When reaching the expansion step in the top-level tree, a second search rooted at the newly expanded node begins instead of a playout or static evaluation. The aggregated results are then propagated up the top-level tree, and the second-level tree is discarded.

Formally, we denote N the total node budget allocated to the search, with N_{top} and N_{sec} being the number of nodes allocated to the top- and second-level trees respectively ($N = N_{\text{top}} + N_{\text{sec}}$). We note P the number of expansions performed during the entire search. Every iteration of a two-level search algorithm performs N_{sec} playouts. Thus, we are able to effectively conduct $P = N_{\text{sec}} \times N_{\text{top}}$ playouts while storing only N nodes in memory. Because of this, two-level search algorithms are often denoted as “squared” variants, due to the total node budget N being equally divided between the top- and second-level trees in most cases ($N_{\text{top}} = N_{\text{sec}}$).

This method is easily generalizable to any MCTS variant, as it simply performs a second search with a different root, and is an efficient way to reduce memory usage by a factor of $\frac{N}{N_{\text{top}} \times N_{\text{sec}}}$ compared to a single-level search. However, efficacy may still be reduced due to discarding fine-grained information gathered in second-level trees. For clarity, we use the “squared” suffix to differentiate base algorithms from their two-level counterparts—e.g. UCT vs UCT^2.

Even though the intermediary steps that led to a specific decision are discarded, human players arguably keep a general sense of which steps led to better outcomes. For example, if a particular action proved to be good in a subcalculation, one can infer that said move is of interest to other subproblems. This form of online learning is emulated by variants of MCTS incorporating AMAF heuristics (see Section 2.2).

Incorporating AMAF heuristics within a two-level search allows AMAF values in top-level tree to direct subsequent second-level searches. While HRAVE might be sensible from a cognitive standpoint—since it stores information learned online for the whole tree—the cognitive plausibility of GRAVE and RAVE is more debatable due to their requirements of storing multiple, separated heuristics along each node.

Another critique of the use of AMAF values is that they store heuristics about all moves played, no matter how inconsequential they might be to the overall search. A more

accurate implementation might only retain information about a few "important" moves, as humans tend to react more strongly to unexpected stimuli. In the context of MCTS, this can translate to actions whose recorded value highly differs from the mean value of their parent node. While not explored within this thesis, we propose to constrain AMAF tables by additionally storing the mean deviation from expected node rewards. Upon backpropagation, the algorithm computes the difference between the expected reward of the node n' reached by playing action a, then computes its absolute difference with the expected reward of the parent node n. If this absolute difference is higher than the lowest one stored in the AMAF table, then this least surprising entry is replaced.

6.1.2. Node recycling

Node recycling (Powley et al., 2017) is a modification of MCTS algorithms allowing them to reuse the space taken by less important nodes to continuously expand the search tree even when under a fixed node budget.

It works by allocating a node pool of size N, filled with empty entries as the beginning of the search. While at least one empty entry is available, newly expanded nodes can be stored in the pool without issue. However, if no empty entry is available, one of the filled entries is recycled by replacing its contents with the newly expanded node. Nodes to recycle are chosen based on the best-first property of MCTS algorithms: branches that are deemed promising are visited more often. By adding a Least Recently Used (LRU) cache to the node pool, removing nodes as we visit them and pushing them to the back of the LRU as we exit them, we can assume that the node at the front of the LRU cache is an unpromising node. This makes it a good candidate for recycling while limiting the effect on the MCTS algorithm's performance. The ordering of removal and push-back operations while traversing the tree ensures that only leaf nodes can ever appear at the front of the LRU, so that no interior node would be recycled.

Using a first-child right-sibling representation for the search tree, along with an intrusive linked list representing the LRU cache, allows the implementation of node recycling with no added time complexity, along with a constant factor to the memory footprint of individual nodes (Figure 8).

Node recycling functions similarly to the Atkinson-Shiffrin model of working memory (Atkinson & Shiffrin, 1968), notably on the key aspect of repetition cementing the persistence of stored information. When combined with RAVE algorithms, it also acts as a light selectivity operator: unlike UCT, the AMAF selection strategy does not always expand recycled children in the root node. Thus, recycled nodes may not be reexpanded again if they were deemed too unfavorable.

We denote node recycling variants of algorithms with a lowercase "r" for "recycling", e.g. UCTr. Recycling schemes can be combined with two-level search by allowing nodes to be recycled within the top-level, second-level or both trees. In this case, both suffixes are used, e.g. HRAVEr^2.

6.2. Selectivity

Human players were observed to be highly selective in the moves they decide to explore. In the context of game-playing agents, this is often referred to as *pruning*. Some algorithms, like AlphaBeta (Knuth & Moore, 1975), prune branches that are proven to be unable to change the result at the root node. Other approaches rely on heuristics to decide which

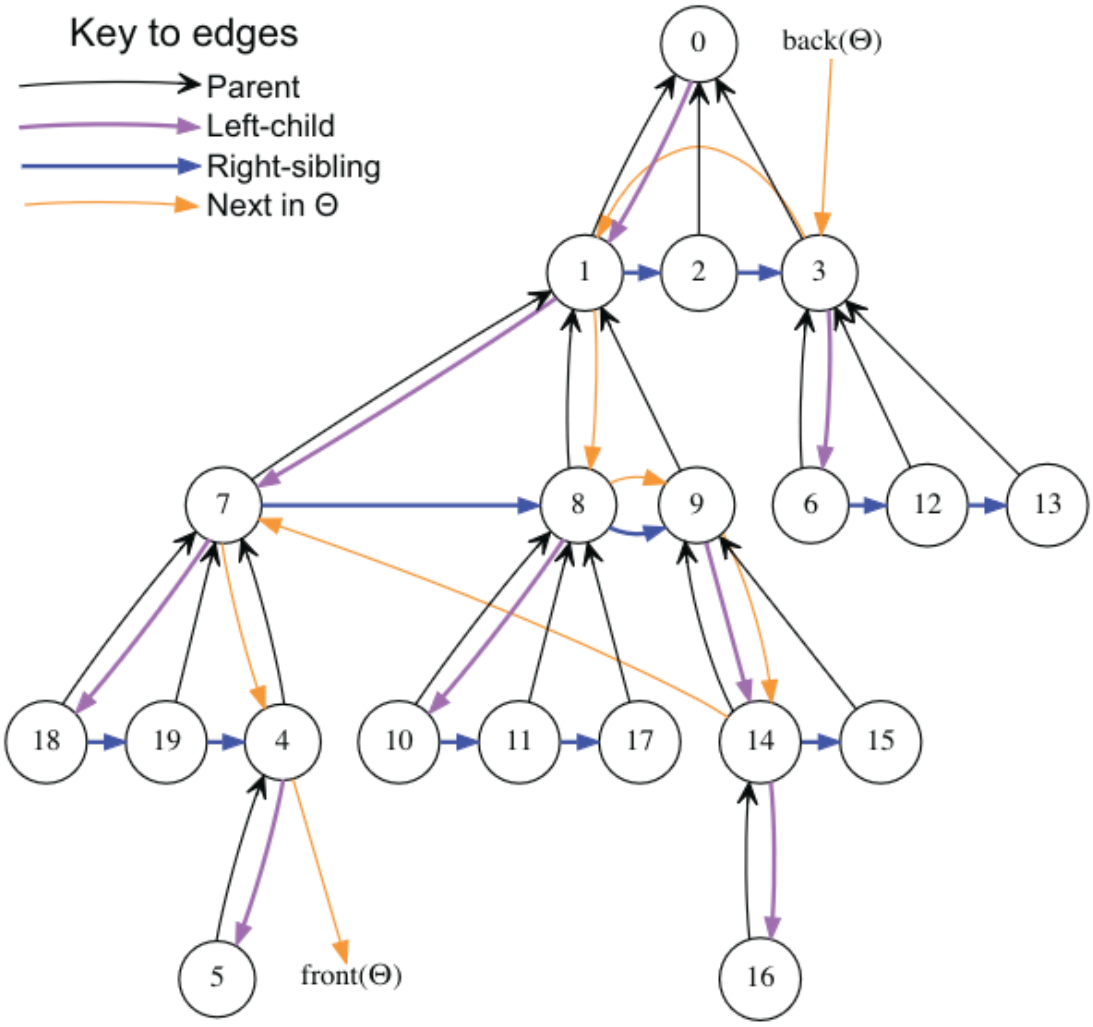


Figure 8: Synthetic tree structure with integrated node recycling (Powley et al., 2017). Θ represents the LRU cache's intrusive doubly-linked list. "Previous node" pointers are omitted for clarity.

branches should be pruned in each node, for example, by pruning branches evaluated under a given threshold by a neural network (Moriarty & Miikkulainen, 1994) (Figure 9). These latter approaches are reminiscent of the DPTC (Kahneman, 2011), using an intuitive system to guide the search routine.

Both methods are directly applicable to MCTS algorithms, provided we have access to heuristics providing a policy over the action set. In cases where domain-specific knowledge is not readily available, such as GGP (Genesereth et al., 2005) systems, this is usually difficult—although some methods have been devised (Soemers et al., 2022; Soemers et al., 2023; Stephenson et al., 2021). The rest of this section will focus on methods applicable without prior domain knowledge or general policy providers, although their inclusion is discussed when applicable.

A known method that can work as a pruning-like mechanism for MCTS is Sequential Halving (SH) (Karnin et al., 2013). SH works by dividing a known budget of iterations between k actions and $\log_2 k$ rounds. After each round, actions are ordered based on their sampled expected reward, and the lower half is pruned before moving to the next

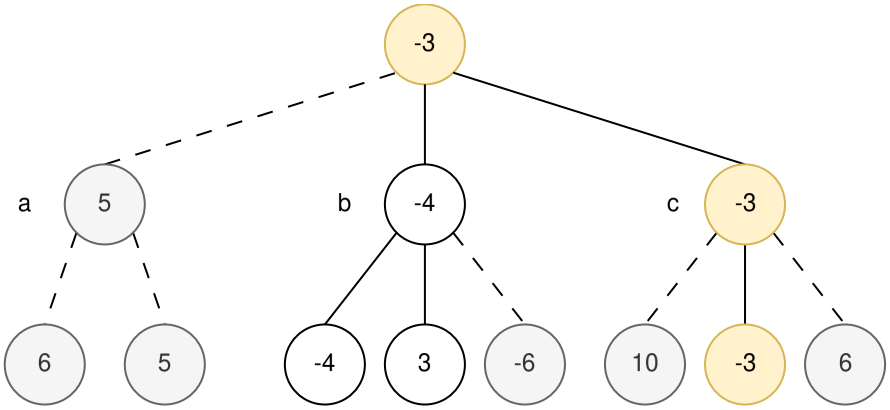


Figure 9: Example of a tree pruned using heuristics to decide which branches to keep in (Moriarty & Miikkulainen, 1994). Pruned branches and states are greyed, resulting in move c seeming to be the best choice when a full-width search would have selected a.

round. When a single action remains, the algorithm stops. SH was originally designed to minimize simple regret rather than cumulative regret in MAB problems (Figure 4). Simple regret ignores the cost of repeatedly sampling arms and only measures the regret of the final decision. It has the advantage of better matching the intuition behind tree search: simulations made while thinking do not matter as long as the played move is the best one. SH has been successfully used as a selection policy within MCTS, either only at the root node or within the whole tree (Cazenave, 2015). It has also been extended to support scores supplied by heuristic functions like AMAF values or neural networks (Fabiano & Cazenave, 2021).

Figure 4: Definitions of the various notions of regret.

- Cumulative regret: $\mathcal{R}_{\text{cum}} = \sum_{r \in \texttt{round}} (p_{i^*} - p_{\hat{r}})$
- Simple regret: $\mathcal{R} = p_{i^*} - p_{\hat{\imath}}$
- Simple regret 0-1: $\mathcal{R}_{\text{0-1}} = \mathbb{I}_{i^* = \hat{\imath}}$

With p_i the expected reward of arm i, i^* the arm with the highest expected reward, $\hat{r}$ the arm chosen by the selection policy at round r and $\hat{\imath}$ the arm chosen after the last round.

The main drawback of SH—requiring a fixed budget—has been addressed by the Anytime SH variant (Sagers et al., 2024). This algorithm works by increasing the number of iterations allocated to each arm exponentially, while still pruning half of the action space after each round. Once only one arm remains, if the time budget allocated to the search is not exhausted, the algorithm restarts while keeping all the information gathered about each arm in previous rounds. It has been empirically shown to be competitive with SH in minimizing simple regret and in playing strength on ten board games.

The main issue of SH and its anytime variant as cognitively plausible models of selectivity is their pruning of a fixed percentage of arms, regardless of the distribution of expected rewards among them. While no formal study of how humans approach such a situation has been conducted to our knowledge, we hypothesize that selectivity depends on the uncertainty of expected outcomes.

$$I(x) = -\log_2 p(x) \tag{12}$$

Following this intuition, we propose **Entropy Pruning (EP)**, an anytime selection strategy partitioning the action space based on the entropy of expected rewards. EP works by

sampling actions in A in a similar way to Anytime SH. Once each arm has been sampled, the action set is partitioned based on the entropy $H(A^i)$ (Equation 5) of the observed rewards of arms in A^i and the information $I(a)$ (Equation 12) of each arm $a \in A^i$. In order for $H(A^i)$ and $I(a)$ to be well-defined, we derive a probability distribution over A^i by using the softmax transformation on the observed rewards of each arm (Equation 13). At the end of round i, we define $A^{i+1} = \{a \mid I(a) \leq H(A^i), a \in A^i\}$, with ε being a parameter controlling the sharpness of the decision boundary.

$$p(a) = \frac{\exp s(a)}{\sum_{b \in A} \exp s(b)} \quad \text{with } s(a) \text{ the score associated to action } a \in A \tag{13}$$

In order to maintain an anytime property, EP also incorporates a rollback mechanism. When $c \log_2 |A^i| \leq H(A^i)$—with $c \in [0, 1]$ controlling the sharpness of the decision boundary—we return to the previous action set A^{i-1}. As the number of overall samples collected grows, the decision boundary should become sharper to avoid stunting the algorithm. Models of how confidence in sampled values evolves for human players calculating lines during play do not exist to our knowledge. Instead, we use the Drift Diffusion Model (DDM) cumulative (Bogacz et al., 2010; Ratcliff et al., 2016) as a parametrized easing function for c based on the number of samples collected by EP (Equation 14). The DDM cumulative describes the process and dynamics of decision-making under the collection of noisy evidence, which maps nicely to the way MCTS operates. We define $c_i = \mathrm{DDM}_{k_c}(N)$ the value of c at round i, with N the total number of collected samples. The k_c parameter can be intuitively understood as controlling how "confident" an agent tends to be. In the specific case of $k_c = 0$, pruning never occurs.

$$\mathrm{DDM}_k(n) = 1 - e^{-kn} \text{ , with } k \text{ controlling the value at which } \mathrm{DDM}(n) = 0.5 \tag{14}$$

To account for the difference in samples inherent to previously pruned branches being reintroduced, we attempt to use UCB1 scores instead of raw expected rewards (Figure 10). This does not improve or decrease the performance of the algorithm in any significant way. Overall, EP performs slightly worse than SH and Anytime SH at minimizing $\mathcal{R}$, but matches their performance for $\mathcal{R}_{0\text{-}1}$ in synthetic MAB problems.

Additionally, we observed the effect of varying k_c both on $\mathcal{R}$ and $\mathcal{R}_{0\text{-}1}$ (Figure 11). Higher and lower values of k_c tend to "stunt" the identification of the best arm, since the decision boundary for rollbacks happens too early for any evidence to have been collected, or too late to make meaningful progress. The value of 0.01 is the most efficient at minimizing both regrets.

Prior scores obtained from heuristic functions like neural networks can be integrated into EP by performing a first pruning step before any sampling takes place. To mimic the way human players completely ignore "intuitively bad" moves, we can force the algorithm to never rollback on this early, prior-informed decision.

6.3. Bounded rationality

Methods modelling the notion of satisficing in game-playing agents are varied. (He et al., 2024) incorporates this notion through a top-2 difference mechanism: given a policy, if the difference in score between the top-2 options is greater than a fixed threshold, search is skipped entirely and the best option is returned instead. While motivated by the DPTC

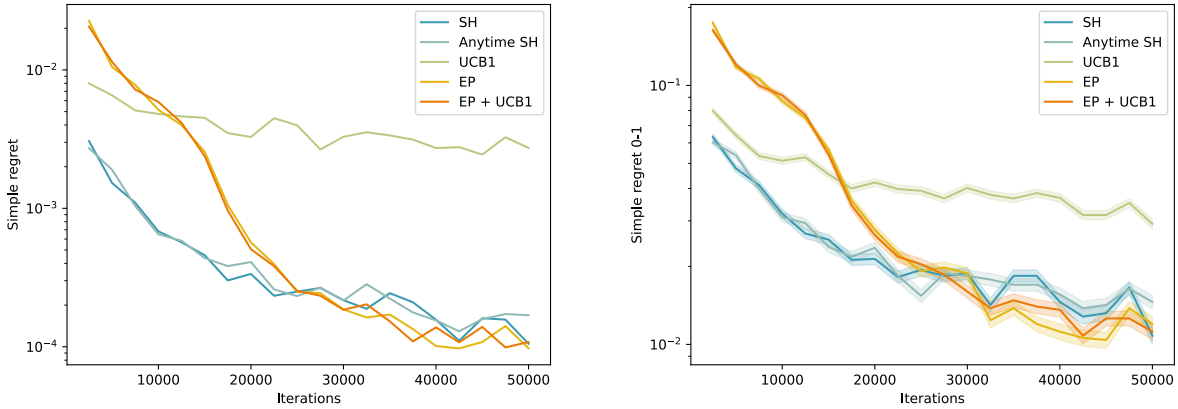


*Figure 10: $\mathcal{R}$ **(left)** and $\mathcal{R}_{0\text{-}1}$ **(right)** for EP ($k_c = 0.05$, using either raw expected rewards or UCB1 with $c = \sqrt{2}^{-1}$) compared to Anytime SH, SH and UCB1 with $c = \sqrt{2}^{-1}$ measured on 5000 synthetic 50-armed bandit problems random rewards following $\mathcal{N}(\mu, 1)$, with $\mu \sim \mathcal{N}(0, 1)$ for each arm. We use a budget of iterations to smooth differences in implementation efficiency.*

(Kahneman, 2011) rather than satisficing, this approach has also been considered a cognitively plausible model of the latter (Schwartz et al., 2002; Vickers, 1979). We refer to this approach as **gap-based aspiration**. KL-regularized search (Jacob et al., 2022) has also been applied to human-like game-playing agents in an effort to combine the advantages of searching and behaviorally aligned policies. The key idea is to measure the divergence between the policy obtained through search and the initial policy obtained from a neural network trained through imitation learning on human players. This allows the search to refine initial judgments without straying too far from a behaviorally aligned policy. KL-regularized search frames the "effort" to change the original policy as an alternative notion of regret, which is supported by information-theoretic models of bounded rationality (Braun & Ortega, 2014).

Based on this prior work, we motivate the cognitive plausibility of EP as a selection strategy by implementing gap-based aspiration. If

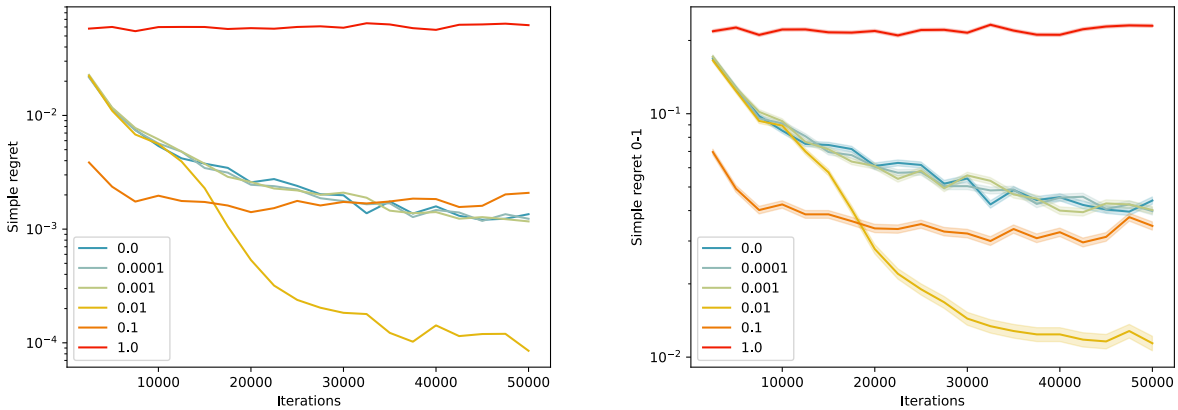


*Figure 11: $\mathcal{R}$ **(left)** and $\mathcal{R}_{0\text{-}1}$ **(right)** of EP with varying values of k_c. The measures are averaged over 5000 synthetic 50-armed bandit problems with random rewards following $\mathcal{N}(\mu, 1)$, with $\mu \sim \mathcal{N}(0, 1)$ for each arm.*

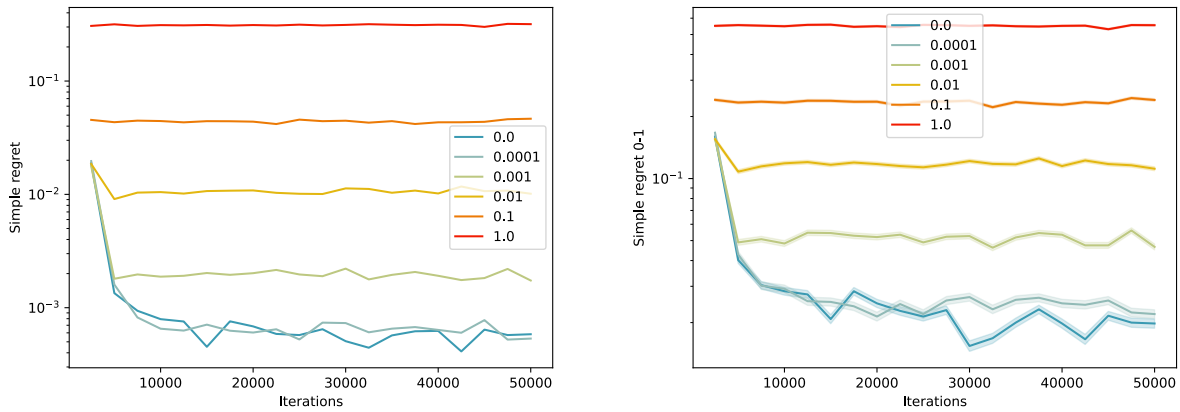


*Figure 12: $\mathcal{R}$ **(left)** and $\mathcal{R}_{0\text{-}1}$ **(right)** of EP with varying values of k_s. The measures are averaged over 5000 synthetic 50-armed bandit problems with random rewards following $\mathcal{N}(\mu, 1)$, with $\mu \sim \mathcal{N}(0, 1)$ for each arm.*

$\text{softmax}(r(a_{\text{top1}})) - \text{softmax}(r(a_{\text{top2}})) \geq s$, where r is the expected reward of action a_i, we reach a satisficing condition and return a_{top1} as the selected action.

As with the notion of doubt introduced to implement the selective aspect of EP, it is natural for s to evolve as more evidence is collected. We use the same DDM cumulative model (Equation 14) to define $s_i = 1 - \text{DDM}_{k_s}(N)$, so that s decreases with the average number of samples per action. The k_s parameter intuitively controls how "lazy" an agent is in searching for the optimal arm. $k_s = 0$ disables satisficing entirely.

We observe the behavior of an agent using EP as a selection strategy on synthetic MAB problems when varying the k_s parameter in Figure 12. The agent stops searching earlier when k_s is high, with clear lower bounds being reached in both $\mathcal{R}$ and $\mathcal{R}_{0\text{-}1}$. "Lazy" agents are less efficient at predicting the best arm. We do however note that agents implementing satisficing with low k_s values are competitive with the unbounded case, even though we verified that they do stop searching earlier.

Similarly to the inclusion of priors in the selectivity of EP, we can check the satisficing condition before performing any sampling if heuristic scores are assigned to moves. In cases where the top-2 gap is high enough, the search can be skipped altogether.

6.4. Experiments on the game of Go

Our experiments measure the relative playing strength of two-level search, node recycling and pruning mechanisms on the game of Atari Go, which uses a 9×9 board. We focus on assessing that playing strength does not significantly degrade when using cognitively motivated algorithms.

The GRAVE and HRAVE algorithms are first tuned against UCT with an exploration constant of 0.7 to determine suitable values for the bias and, in the case of GRAVE, reference threshold parameters. These parameters were fixed to 10^-2 and 25 respectively for all subsequent experiments.

Playouts are guided by MAST (Finnsson & Björnsson, 2010b) using an ε-greedy sampling strategy with $\varepsilon = 0.4$. MAST statistics are decayed by a factor of 0.2 between successive turns within the same game.

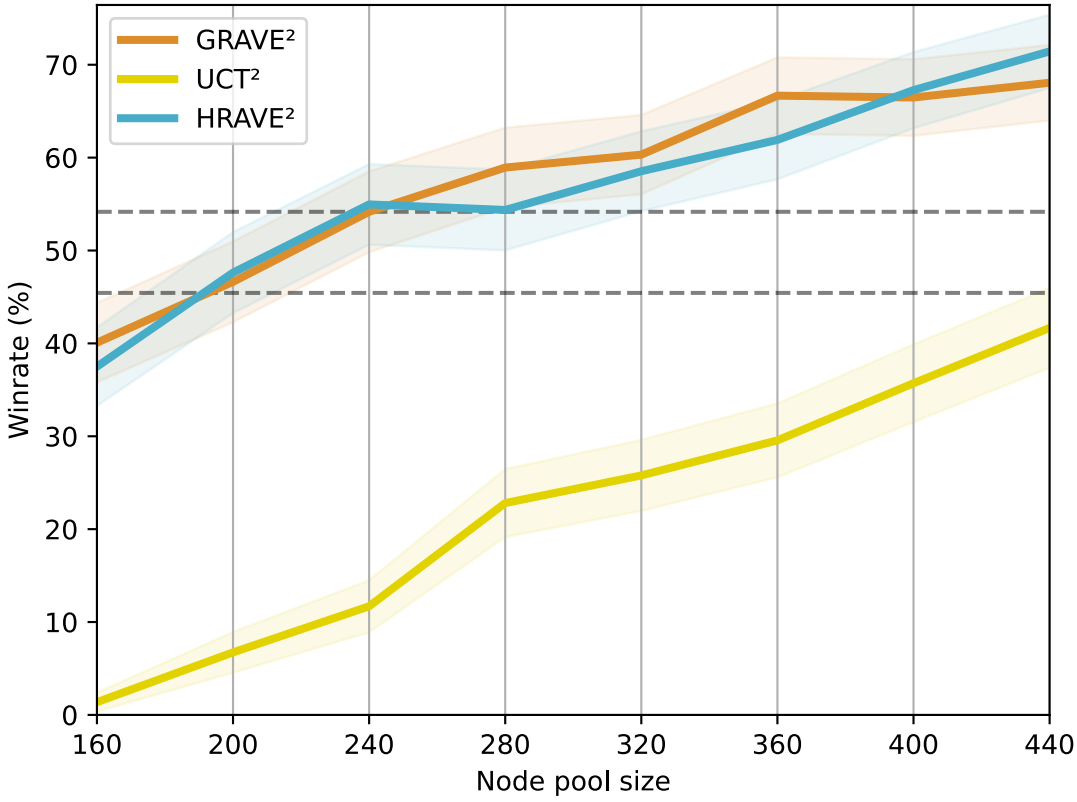


Figure 13: Winrates of two-level search algorithms against GRAVE with $P = N = 10,000$. The dotted lines indicate the 95% confidence interval of the winrate of GRAVE against itself.

6.4.1. Memory-constrained search

We evaluate the relative playing strength of the GRAVE, HRAVE and UCT with two-level search and/or node recycling, using GRAVE with $P = N = 10,000$ as a baseline. Each data point is derived from the results of 250 games with each color. Win rates are reported together with Agresti-Coull confidence intervals.

We first evaluate the impact of two-level search on playing strength when reducing the node budget of the algorithms. The relative win rates are reported as a function of the number of nodes stored by the two-level search algorithms in Figure 13.

GRAVE² achieves comparable playing strength starting from a total node pool size of 200 ($N_\top = N_{\text{sec}} = 100$, corresponding to $P = 10,000$). UCT² requires more than 440 nodes to reach similar results, performing more than four times as many playouts as GRAVE ($P \geq 48\{,\}400$). However, storing AMAF statistics in each node results in a substantially larger memory footprint. GRAVE has an asymptotic memory footprint of $\mathcal{O}(N \times |A|)$, with N the number of nodes stored and $|A|$ the number of legal actions in any position, while UCT's memory footprint is of the order $\mathcal{O}(N)$. HRAVE² performs similarly to GRAVE² while having the same asymptotic memory footprint as UCT since a single AMAF table is stored for the entire tree.

We then turn to node recycling, with a similar experimental setup where the maximum node pool size varies, but the number of playouts performed is fixed to $P = 10,000$. The results reported in Figure 14 show that this approach is less effective at reducing the memory footprint of the search tree while maintaining playing strength. GRAVEr

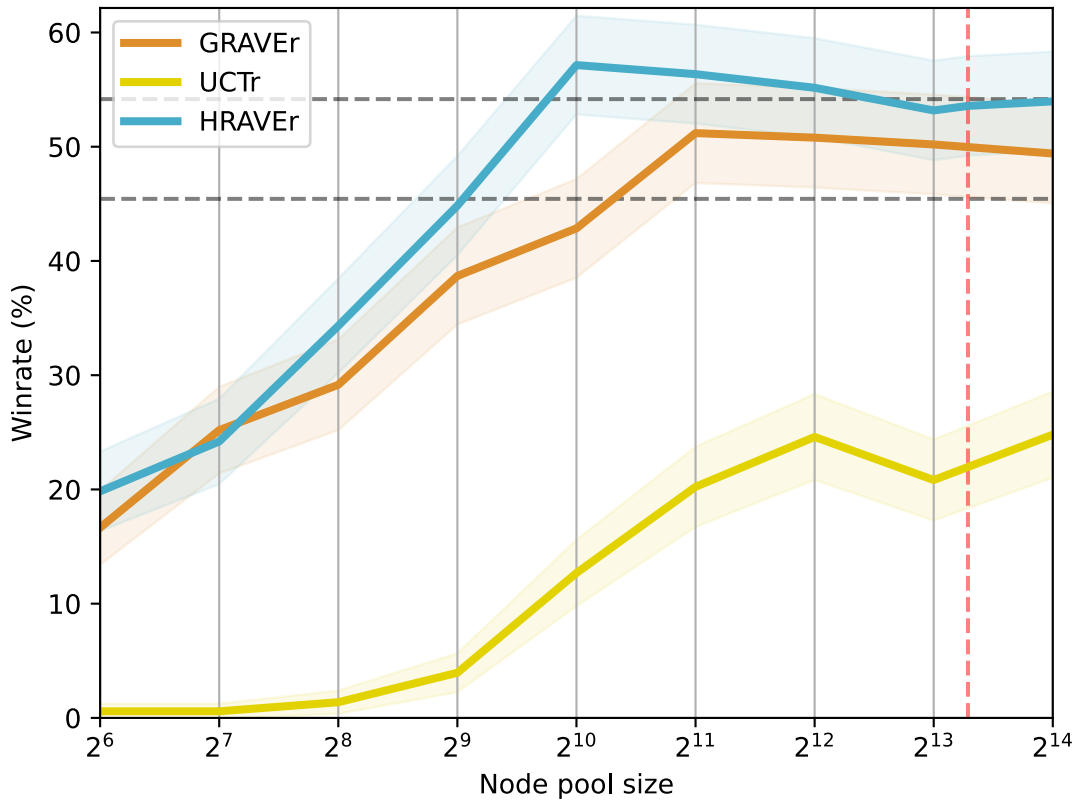


Figure 14: Winrates of GRAVE, HRAVE and UCT with node recycling against GRAVE with $P = 10,000$. The node pool size used is presented on a logarithmic scale. The red dashed line indicates the threshold beyond which node recycling no longer takes effect.

reaches equal performance to GRAVE at $N \approx 1,024$, while HRAVEr reached similar playing strength from $N \approx 512$ onwards.

Finally, we evaluate the performance of combining node recycling with two-level search through two experiments using node recycling in the top- and second-level tree (Figure 15) by varying the number of additional playouts as a ratio r of the node pool size ($P = r \times N$). A ratio of 1.0 corresponds to disabling node recycling entirely.

We are able to reduce the node pool size of GRAVEr² to 160 nodes while maintaining performance comparable to GRAVE. This threshold is noteworthy because it prevents the tree from expanding all root children while still reaching depth ≥ 1, indicative of a form of pruning. HRAVEr² further improves this result, reaching a playing strength equal to GRAVE with only 120 nodes.

Applying node recycling to the second-level tree does not improve the playing strength of either algorithm significantly. One possible explanation is that the top-level tree may not accumulate sufficient information to compensate for the loss of finer-grained statistics in the second-level tree, thus wasting the additional playouts.

6.4.2. Entropy Pruning

We now evaluate the performance of using EP as a selection strategy. Interior nodes of the tree use the UCT selection strategy with $c = \sqrt{2}$.

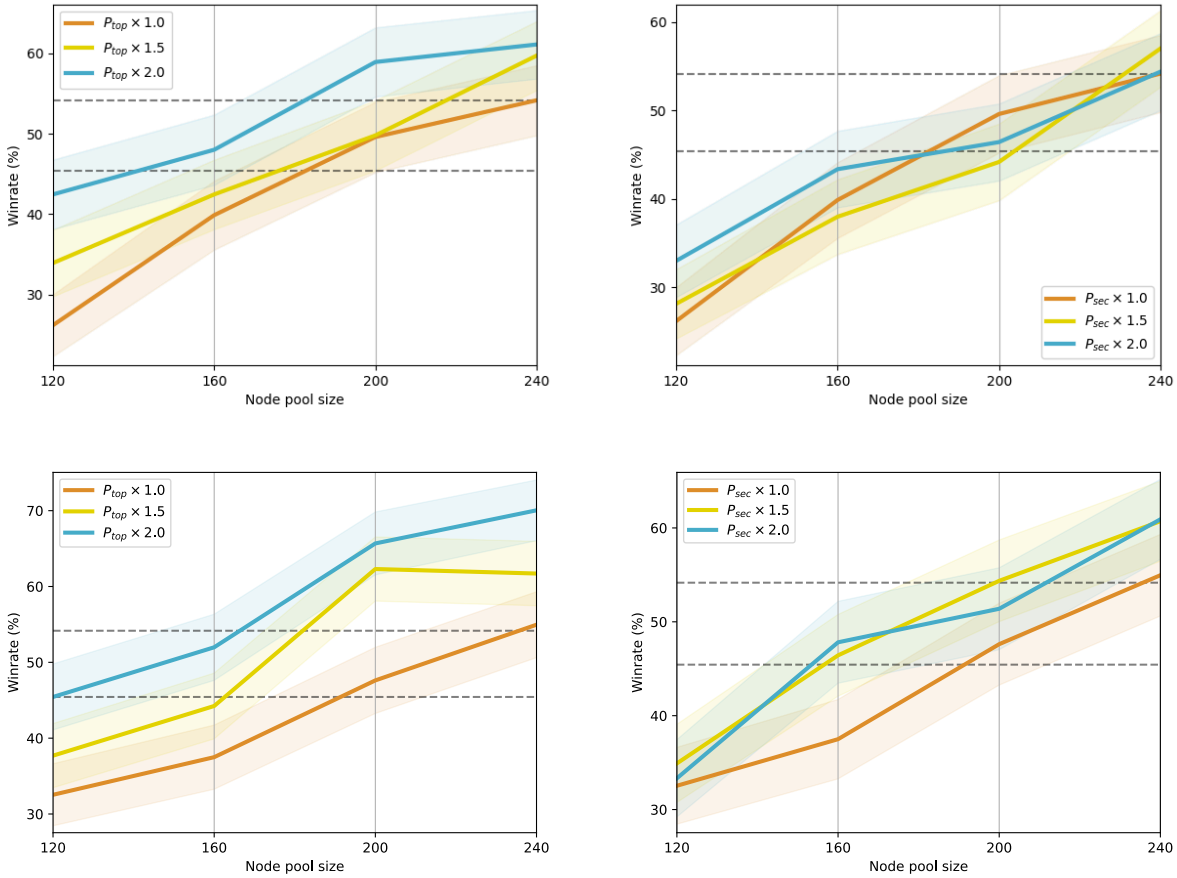


*Figure 15: Winrate of GRAVEr² (**top**) and HRAVEr² (**bottom**) against GRAVE ($P = 10,000$). We vary the ratio of playouts to stored nodes in the top- (**left**) and second-level tree (**right**).*

First, we measure EP's playing strength relative to a pure UCT search, meaning all nodes use UCT as their selection strategy with $c = 0.7$. We vary both its confidence k_c and laziness k_s (Table 4). The impact of both parameters follows the same tendencies as in synthetic MAB problems. However, the improved $\mathcal{R}$ and $\mathcal{R}_{0\text{-}1}$ observed over UCB1 do not translate to an increase in winrate over UCT in Go.

Using the best parameters from these experiments ($k_c = 0.1$, $k_s = 0$), we compare the playing strength of SH, Anytime SH and EP as root strategies (Figure 16). SH surpasses all other approaches in this scenario, being the only root selection strategy to consistently surpass raw UCT. EP is comparable to Anytime SH, although it performs slightly worse

k_c	Winrate	k_s	Winrate
0.0	$40.9 \pm 6.0\%$	0.0	$49.2 \pm 6.2\%$
0.01	$40.2 \pm 6.0\%$	0.01	$31.1 \pm 5.7\%$
0.05	$43.3 \pm 6.1\%$	0.05	$22.8 \pm 5.2\%$
0.1	$48.4 \pm 6.1\%$	0.1	$9.4 \pm 3.6\%$
0.5	$41.3 \pm 6.1\%$	0.5	$3.1 \pm 2.1\%$
1.0	$39.4 \pm 6.0\%$	1.0	$3.1 \pm 2.1\%$

*Table 4: Winrate of UCT with $c = 0.7$ using EP as a root selection strategy against UCT with the same c used in all nodes. Both algorithms are given $5,000$ iterations per move. We vary the confidence k_c (**left**) and satisficing k_s (**right**) of EP between 0 and 1.*

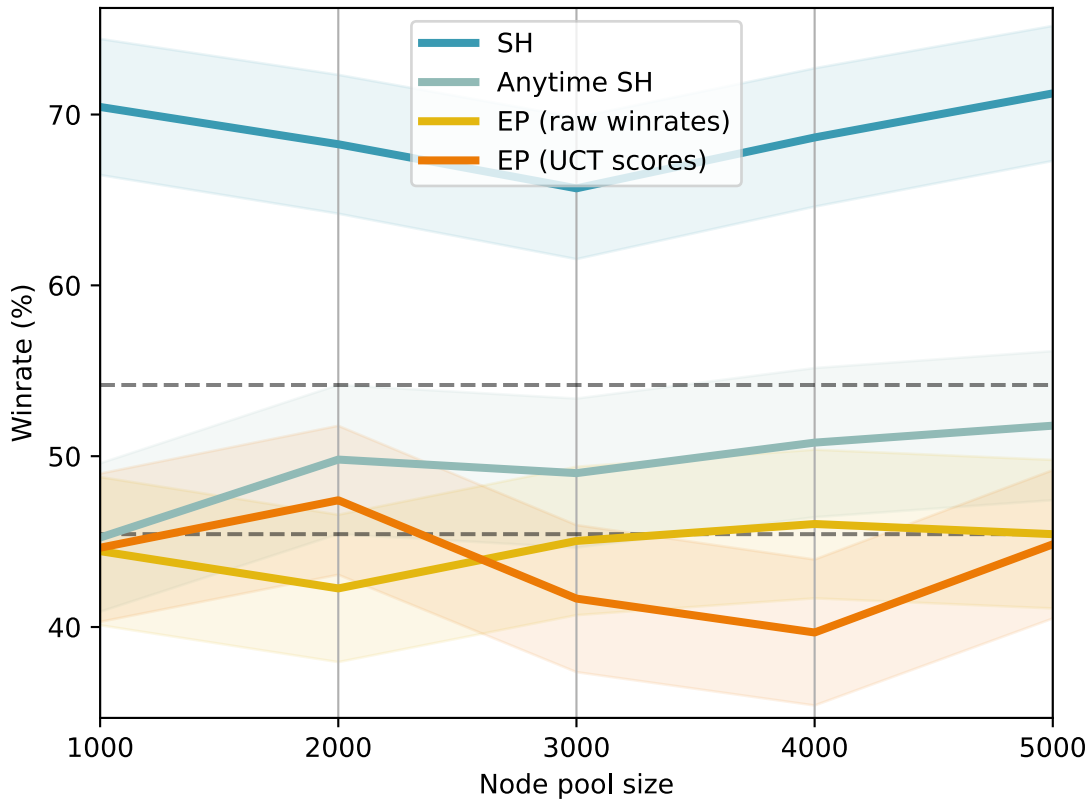


Figure 16: Winrate of UCT with $c = 0.7$ using EP, SH and Anytime SH as root selection strategies against UCT with the same c used in all nodes. EP uses the best performing parameters derived in Table 4.

overall. The use of UCT scores instead of raw winrates in EP does not seem to alter its performance significantly.

The program used to run these experiments can be found at https://codeberg.org/thinkingrocks/gomuki.

6.5. Discussion

Our experiments show that designing agents more closely aligned with human cognition is feasible while maintaining the overall playing strength of classic algorithms. We manage to reach equal playing strength to GRAVE while reducing the node pool size down to around 1.2%. While a long shot from realistic estimates of human memory capacity, this is an encouraging result. Incorporating heuristics may enable us to further reduce the number of nodes stored while maintaining acceptable playing strength. Currently, most of our experiments rely on random playouts, which are both noisy and unrealistic for human players to employ.

Using EP as a model of bounded rationality and selectivity yields unsatisfactory results in playing strength. However, the introduction of parameters related to satisficing and confidence may allow a better modelling power. Further work is needed to both validate the model's cognitive plausibility through experiments and study its behavior when paired with heuristic functions.

The gap-based approach used to incorporate elements of satisficing and bounded rationality in EP is only an example among many competing models (Braun & Ortega, 2014; Jacob et al., 2022). We identify bounded rationality as a key feature of cognitively plausible search systems, especially when extended to account for factors outside of pure score-related considerations (e.g. accounting for time constraints, patience, or focus).

While this section focused solely on MCTS variants, we plan to study depth-first search algorithms (e.g. AlphaBeta (Knuth & Moore, 1975)) as cogntivively plausible search methods. In games such as Bao or Awele, where search plays a more important role in human planning, such algorithms might capture human play more accurately. Depth-first algorithms model the sequence-oriented way human players search, rather than the repeated, mean-reward sampling performed by MCTS algorithms. While modifications must be made to account for bounded rationality and selectivity, memory constraints can be handled simply by limiting the depth of search.

7

Discussion and future work

This thesis has presented the first classification of human-like game agents beyond the functionalist view prevalent in modern literature. We identified three complementary axes upon which to evaluate and classify such agents: their **generalizability**, how well can they transfer knowledge across games; their **qualities**, being either behavioural alignement, cognitive plausibility or both; and their **granularity**, how much of the range of individual human behaviors can they capture. We further refine the idea of behavioral similarity beyond move-level alignment, acknowledging that human players think in terms of sequences of moves, and cementing games as a social and interactive activity. Building upon this classification, we focused on the exploration of cognitively plausible algorithms for two major aspects of game playing agents: **search** through two-level search, node recycling and the EP selection strategy; and **heuristics** with a Chunked GNN approach applied to Chess.

This work also reviews and introduces methodologies for measuring behavioral alignment within domains formalized as MDPs. We notably identify the most commonly used metric, move matching, as unsuitable for measuring alignment, and propose alternative formulations for **one-to-one** divergence. Two additionnal ways to define alignement are proposed: **one-to-many** for assessing the likelihood that an observed behavior is distinguishable from a population of behaviors; and **many-to-many** to measure how well a population of behaviors captures the distribution and variety of another target population.

Overall, extending research on human-like agents beyond behavioral alignment opens a wide array of questions for both AI and cognitive science researchers. In a continuation of their role as test benches in the pursuit of superhuman intelligence, games provide a perfect mix of complexity, variety and formalism to explore these prospects before hopefully generalizing them to other fields.

7.1. Future work

The results and theoretical considerations formulated in this exploratory work are meant to form a foundation upon which to build a broader field of research for human-like agents, both in games and other applicable domains. While this work highlights many potential research directions, it was able to explore only a few of them at a surface level. This final section presents some of the planned future work stemming from this thesis.

7.1.1. Relation between cognitive plausibility and behavioral alignment

One of the most important questions left open is the link between cognitively plausible algorithms and behavioral alignment. None of the approaches presented in this manuscript were assessed on their impact regarding behavioral alignment. We plan to merge these approaches to search and heuristics into a cohesive cognitively-plausible model, which will be tested and adapted for behavioral alignment (Rautureau & Piette, 2025).

7.1.2. Search algorithms

While bounded rationality and selectivity were identified as core components of how humans plan in games, the EP selection strategy presented in Chapter 6 has only been studied on aspects of playing strength, while its capacity to model a range of behaviors has not been assessed. We plan to conduct a deeper study of this aspect of human-like planning in games by assessing the impact of different parameters of EP on behavioral alignment when paired with a heuristic function, along with further refinements to the model informed by previous work on bounded rationality. Additionally, investigating AlphaBeta as a cognitively-plausible search algorithm, or more cognitively-accurate caching of search information, are secondary projects related to search. The modeling of pondering time during search is also a key factor on how humans are able to distinguish artificial agents (Laarhoven & Ponukumati, 2023; Zhang et al., 2024), and should be explored further.

7.1.3. Interactive and social aspects of games

Social and interactive aspects of behavior and cognition are absent from this thesis, although they form a key aspect of games. Game-playing agents typically model their opponent as a mirror image of themselves, with access to the same knowledge, reasoning and goals. In contrast, humans model their opponent based on their perceived skill, motivations and preferences (Thagard, 1992). We aim to study and incorporate this aspect within our models through the perspective of Theory of Mind (ToM), and building upon previous work done on Opponent Modelling in games (Carmel & Markovitch, 1993; Donkers et al., 2001; Jansen, 1993). SoM measures of behavioral alignment are of peculiar importance to this effort as they capture interactive information. We thus plan to research and develop such methods in parallel to our exploration of interactive behavior.

We identify *cheating* as a central modelling problem related to the material aspect of game playing. Computer programs model games as abstract problems, and granting them the ability to thwart said abstraction in order to gain a strategic advantage has not been extensively studied. One reason is that the symbolic and abstract nature of computerized games makes it difficult to accurately reflect material aspects. Studies of this aspect of game playing are currently being conducted by the Ludus Ex Machina project, with promising results.

7.1.4. Generalizable heuristics

Generalizability is a less extensively studied aspect of machine learning in games, although some work has been done on the matter within the field of GGP (Soemers et al., 2023; Stephenson et al., 2021). We plan to apply GNN architectures (Scarselli et al., 2008) to encompass multiple related games at once (Chess, Shogi, Breakthrough and Hnefatafl) while using a single network. The symbolic nature of game states makes their representation as graphs natural (e.g. graphs are the foundation of state representations in Ludii (Piette et al., 2020)). Learning heuristics on these graphs might allow them to be more

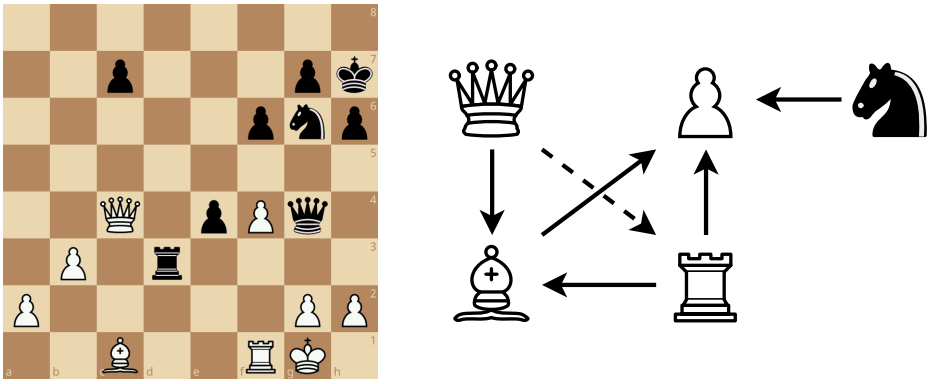


Figure 17: Example pattern used by MORPH. A subgraph of the position is isolated, with dashed and solid lines indicating discovered and direct attacks respectively.

easily generalized, as the input domain of the neural network would remain the same across games. Furthermore, graph-like patterns have already been applied as cognitively plausible representations in games like Chess (Levinson, 1994; Levinson & Snyder, 1991) (Figure 17) and Connect-4 (Mandziuk, 2011). The applicability of graph-based heuristics in game were graph-like patterns are less naturally derived (e.g. Go, Othello) is also of interest to assess the generalizability of this approach beyond Chess-like games.

7.1.5. Beyond perfect information games

While this manuscript focuses exclusively on perfect-information games, chance and partial information are important aspects of tabletop games as a whole, with most casual games including some or both of them. They are also particularly relevant to human-like play, as our appraisal of statistics is riddled with biases (Tversky et al., 1982). The application of models studying these biases (Quiggin, 1982; 1993) has not been studied for game-playing agents, and forms an exciting area of future research.

List of abbreviations

AI	Artificial Intelligence
AMAF	All Moves As First
DDM	Drift Diffusion Model
DPTC	Double Process Theory of Cognition
EP	Entropy Pruning
GGP	General Game Playing
GRAVE	Generalized Rapid Action Value Estimation
LOF	Local Outlier Factor
LRU	Least Recently Used
MAB	Multi-Armed Bandit
MAST	Move Average Sampling Technique
MCTS	Monte-Carlo Tree Search
MDP	Markov Decision Processes
MRI	Magnetic Resonance Imaging
PD	Playstyle Distance
POD	Policy Outlier Detection
PS	Playstyle Similarity
RAVE	Rapid Action Value Estimation
RL	Reinforcement Learning
SH	Sequential Halving
SoM	Sequences of Moves
UCT	Upper Confidence bounds applied to Trees
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction

Bibliography

Allis, L. (1994). *Searching for Solutions in Games and Artificial Intelligence*. Rijksuniversiteit Limburg. https://doi.org/10.26481/dis.19940923la

Anthony, T., Tian, Z., & Barber, D. (2017). Thinking Fast and Slow with Deep Learning and Tree Search. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*.

Atherton, M., Zhuang, J., Bart, W. M., Hu, X., & He, S. (2003). A Functional MRI Study of High-Level Cognition. I. The Game of Chess. *Cognitive Brain Research*, *16*(1), 26–31.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and Its Control Processes. In *Psychology of Learning and Motivation: Vol. 2. Psychology of Learning and Motivation* (pp. 89–195). Elsevier.

Bart, W. (2012). Moves in Mind: The Psychology of Board Games. *Int. J. Gaming Comput. Mediat. Simulations*, *4*(3), 92–94. https://doi.org/10.4018/JGCMS.2012070107

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. de O., Raiman, J., Salimans, T., Schlatter, J., … Zhang, S. (2019). Dota 2 with Large Scale Deep Reinforcement Learning. *Corr*, *abs/1912.06680*.

Bhattacharyya, A. (1946). On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics*, 401–406.

Binet, A. (1894,). *Psychologie des grands calculateurs et joueurs d'échecs*. Paris : Hachette.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The Neural Basis of the Speed–Accuracy Tradeoff. *Trends in Neurosciences*, *33*(1), 10–16.

Braun, D., & Ortega, P. (2014). Information-Theoretic Bounded Rationality and \varepsilon-Optimality. *Entropy*, *16*(8), 4662–4676. https://doi.org/10.3390/e16084662

Breuker, D. M. (1998). *Memory versus Search in Games*.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00*, 93–104. https://doi.org/10.1145/342009.335388

Carmel, D., & Markovitch, S. (1993). *Learning Models of Opponent's Strategy Game Playing*.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S. A., Abbeel, P., & Dragan, A. D. (2019). On the Utility of Learning about Humans for Human-AI Coordination. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

Cazenave, T. (n.d.). *Generalized Rapid Action Value Estimation*.

Cazenave, T. (2015). Sequential Halving Applied to Trees. *IEEE Trans. Comput. Intell. AI Games*, *7*(1), 102–105. https://doi.org/10.1109/TCIAIG.2014.2317737

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The Perceptual Aspect of Skilled Performance in Chess: Evidence from Eye Movements. *Memory & Cognition*, *29*(8), 1146–1152. https://doi.org/10.3758/BF03206384

Chase, W. G., & Simon, H. A. (1973). Perception in Chess. *Cognitive Psychology*, *4*(1), 55–81.

Chen, X., Zhang, D., Zhang, X., Li, Z., Meng, X., He, S., & Hu, X. (2003). A Functional MRI Study of High-Level Cognition: II. The Game of GO. *Cognitive Brain Research*, *16*(1), 32–37.

Chu-Husan, H., Kyota, K., Shi-Jim, Y., & Kokolo, I. (n.d.). *Making KataGo HumanSL More Human-Like for Amateur-Level Play*.

Crist, W. (2019). Playing against Complexity: Board Games as Social Strategy in Bronze Age Cyprus. *Journal of Anthropological Archaeology*, *55*, 101078.

Crist, W., De Voogt, A., & Dunn-Vaturi, A.-E. (2016). Facilitating Interaction: Board Games as Social Lubricants in the Ancient Near East. *Oxford Journal of Archaeology*, *35*(2), 179–196. https://doi.org/10.1111/ojoa.12084

Dasen, V. (2020). Play with fate. *Ancient Magic: Then and Now*, 173–191.

De Groot, A. D. (1946). *Thought and Choice in Chess* (p. 315). Noord-Holl. Uitg.

de Voogt, A. (1995). *Limits of the Mind: Towards a Characterisation of Bao Mastership*.

de Voogt, A. J. (2005). *A Question of Excellence: A Century of African Masters*. Africa World Press.

Donkers, H. H. L. M., Uiterwijk, J. W. H. M., & van den Herik, H. J. (2001). Probabilistic Opponent-Model Search. *Information Sciences, Heuristic Search and Computer Game Playing*, *135*(3), 123–149. https://doi.org/10.1016/S0020-0255(01)00133-5

Duch, W., Oentaryo, R. J., & Pasquier, M. (2008). Cognitive Architectures: Where Do We Go from Here?. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence 2008, Proceedings of the First AGI Conference, AGI 2008, March 1-3, 2008, University of Memphis, Memphis, TN, USA: Artificial General Intelligence 2008, Proceedings of the First AGI Conference, AGI 2008, March 1-3, 2008, University of Memphis, Memphis, TN, USA*.

Fabiano, N., & Cazenave, T. (2021). Sequential Halving Using Scores. In C. Browne, A. Kishimoto, & J. Schaeffer (Eds.), *Advances in Computer Games - 17th International*

Conference, ACG 2021, Virtual Event, November 23-25, 2021, Revised Selected Papers: Advances in Computer Games - 17th International Conference, ACG 2021, Virtual Event, November 23-25, 2021, Revised Selected Papers. https://doi.org/10.1007/978-3-031-11488-5_4

Fechner, G. T. (1948). Elements of Psychophysics, 1860. In *Readings in the History of Psychology: Readings in the History of Psychology* (pp. 206–213). Appleton-Century-Crofts. https://doi.org/10.1037/11304-026

Finnsson, H., & Björnsson, Y. (2010a). Learning Simulation Control in General Game-Playing Agents. In M. Fox & D. Poole (Eds.), *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. https://doi.org/10.1609/AAAI.V24I1.7651

Finnsson, H., & Björnsson, Y. (2010b). Learning Simulation Control in General Game-Playing Agents. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, USA, July 11-15, 2010*, 954–959.

Gelly, S., & Silver, D. (2011). Monte-Carlo Tree Search and Rapid Action Value Estimation in Computer Go. *Artif. Intell.*, *175*(11), 1856–1875. https://doi.org/10.1016/J.ARTINT.2011.03.007

Gelly, S., & Teytaud, O. (2006). *Modification of UCT with Patterns in Monte-Carlo Go*.

Genesereth, M. R., Love, N., & Pell, B. (2005). General Game Playing: Overview of the AAAI Competition. *AI Mag.*, *26*(2), 62–72. https://doi.org/10.1609/AIMAG.V26I2.1813

Gobet, F., & Simon, H. A. (1998). Expert Chess Memory: Revisiting the Chunking Hypothesis. *Memory*, *6*(3), 225–255. https://doi.org/10.1080/741942359

Goodfellow, I. J., Bengio, Y., & Courville, A. C. (2016). *Deep Learning*. MIT Press.

Hay, N., Russell, S., Tolpin, D., & Shimony, S. E. (2014, August). *Selecting Computations: Theory and Applications* (Issue arXiv:1408.2048). arXiv. https://doi.org/10.48550/arXiv.1408.2048

He, T., Liao, L., Cao, Y., Liu, Y., Liu, M., Chen, Z., & Qin, B. (2024). Planning Like Human: A Dual-process Framework for Dialogue Planning. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. https://doi.org/10.18653/V1/2024.ACL-LONG.262

Hellinger, E. (1909). Neue Begründung Der Theorie Quadratischer Formen von Unendlichvielen Veränderlichen. *Journal Für Die Reine Und Angewandte Mathematik*, *1909*(136), 210–271. https://doi.org/10.1515/crll.1909.136.210

Helmbold, D. P., & Parker-Wood, A. (2009). All-Moves-As-First Heuristics in Monte-Carlo Go. In H. R. Arabnia, D. de la Fuente, & J. A. Olivas (Eds.), *Proceedings of the 2009 International Conference on Artificial Intelligence, ICAI 2009, July 13-16, 2009, Las Vegas Nevada, USA, 2 Volumes: Proceedings of the 2009 International Conference on Artificial Intelligence, ICAI 2009, July 13-16, 2009, Las Vegas Nevada, USA, 2 Volumes*.

Jacob, A. P., Wu, D. J., Farina, G., Lerer, A., Hu, H., Bakhtin, A., Andreas, J., & Brown, N. (2022). Modeling Strong and Human-Like Gameplay with KL-Regularized Search. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA: International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*.

Jansen, P. J. (1993). KQKR: Speculatively Thwarting a Human Opponent. *ICGA Journal*, *16*(1), 3–17. https://doi.org/10.3233/ICG-1993-16102

Kahneman, D. (2011). *Thinking, Fast and Slow*. macmillan.

Kahneman, D., & Tversky, A. (2013). Prospect Theory: An Analysis of Decision Under Risk. In *World Scientific Handbook in Financial Economics Series* (Vol. 4, pp. 99–127). WORLD SCIENTIFIC. https://doi.org/10.1142/9789814417358_0006

Kantorovich, L. V. (1939). The mathematical method of production planning and organization. *Management Science*, *6*(4), 363–422.

Karnin, Z. S., Koren, T., & Somekh, O. (2013). Almost Optimal Exploration in Multi-Armed Bandits. *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013, JMLR Workshop and Conference Proceedings*, 1238–1246.

Knuth, D. E., & Moore, R. W. (1975). An Analysis of Alpha-Beta Pruning. *Artif. Intell.*, *6*(4), 293–326. https://doi.org/10.1016/0004-3702(75)90019-3

Kocsis, L., & Szepesvári, C. (2006). Bandit Based Monte-Carlo Planning. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings: Vol. 4212. Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings*. https://doi.org/10.1007/11871842_29

Laarhoven, T., & Ponukumati, A. (2023). Towards Transparent Cheat Detection in Online Chess: An Application of Human and Computer Decision-Making Preferences. In C. Browne, A. Kishimoto, & J. Schaeffer (Eds.), *Computers and Games: Vol. 13865. Computers and Games* (pp. 163–180). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-34017-8_14

Levinson, R. A. (1994). *MORPH II: A Universal Agent: Progress Report and Proposal*. Computer Research Laboratory [University of California, Santa Cruz].

Levinson, R., & Snyder, R. (1991). Adaptive Pattern-Oriented Chess. In *Machine Learning Proceedings 1991: Machine Learning Proceedings 1991* (pp. 85–89). Elsevier.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the Brain. *Nature Reviews Neuroscience*, *21*(6), 335–346.

Lin, C.-C., Chiu, W.-C., & Wu, I.-C. (2021). An Unsupervised Video Game Playstyle Metric via State Discretization. In C. P. de Campos, M. H. Maathuis, & E. Quaeghebeur (Eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021: Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*.

Lin, C.-C., Chiu, W.-C., & Wu, I.-C. (2024, August). *Perceptual Similarity for Measuring Decision-Making Style and Policy Diversity in Games* (Issue arXiv:2408.06051). arXiv. https://doi.org/10.48550/arXiv.2408.06051

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Mandziuk, J. (2011). Towards Cognitively Plausible Game Playing Systems. *IEEE Comput. Intell. Mag.*, *6*(2), 38–51. https://doi.org/10.1109/MCI.2011.940626

McIlroy-Young, R., Sen, S., Kleinberg, J. M., & Anderson, A. (2020). Aligning Superhuman AI with Human Behavior: Chess as a Model System. In R. Gupta, Y. Liu, J. Tang, & B. A. Prakash (Eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020: KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*. https://doi.org/10.1145/3394486.3403219

McIlroy-Young, R., Wang, Y., Sen, S., Kleinberg, J. M., & Anderson, A. (2021). Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*.

McInnes, L., Healy, J., & Melville, J. (2020, September). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (Issue arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Mestre, J. P. (2000). *Transfer of Learning from a Modern Multidisciplinary Perspective*. IAP.

Miller, G. A. (1956). The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Moriarty, D. E., & Miikkulainen, R. (1994). Evolving Neural Networks to Focus Minimax Search. In B. Hayes-Roth & R. E. Korf (Eds.), *Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 2: Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 2*.

Murphy, A. H. (1996). The Finley Affair: A Signal Event in the History of Forecast Verification. *Weather and Forecasting*, *11*(1), 3–20.

Nau, D. S. (1983). Decision Quality As a Function of Search Depth on Game Trees. *J. ACM*, *30*(4), 687–708. https://doi.org/10.1145/2157.322400

Newell, A. (1994). *Unified Theories of Cognition*. Harvard University Press.

Piette, É., Soemers, D. J. N. J., Stephenson, M., Sironi, C. F., Winands, M. H. M., & Browne, C. (2020). Ludii - The Ludemic General Game System. In G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, & J. Lang (Eds.), *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020): ECAI 2020 - 24th*

European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020). https://doi.org/10.3233/FAIA200120

Powley, E., Cowling, P., & Whitehouse, D. (2017). Memory Bounded Monte Carlo Tree Search. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *13*(1), 94–100. https://doi.org/10.1609/aiide.v13i1.12932

Quiggin, J. (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Quiggin, J. (1993). *Generalized Expected Utility Theory*. Springer Netherlands. https://doi.org/10.1007/978-94-011-2182-8

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Rautureau, A., & Piette, É. (2025, July). *CogniPlay: A Work-in-Progress Human-like Model for General Game Playing* (Issue arXiv:2507.05868). arXiv. https://doi.org/10.48550/arXiv.2507.05868

Rautureau, Cazenave, & Piette. (2026). Generalized Rapid Action Value Estimation in Memory-Constrained Environments. *Corr*, *abs/2602.23318*. https://doi.org/10.48550/ARXIV.2602.23318

Rautureau, Piette, Goodman, et al. (2026,). Human-like Game AI Beyond Behavioral Alignment. *Under Review for IEEE Conference on Games (Cog)*.

Rautureau, Piette, Penn, et al. (2026). Game AI for Cultural Heritage: Outcomes from the GameTable Third Grant Period Opening Meeting. *ICGA Journal: The Journal of the Computer Games Community*, 13896911261446593. https://doi.org/10.1177/13896911261446593

Reingold, E. M., & Charness, N. (2005). Perception in Chess: Evidence from Eye Movements. *Cognitive Processes in Eye Guidance*, 325–354.

Reitman, J. S. (1976). Skilled Perception in Go: Deducing Memory Structures from Inter-Response Times. *Cognitive Psychology*, *8*(3), 336–356. https://doi.org/10.1016/0010-0285(76)90011-6

Sagers, D., Winands, M. H. M., & Soemers, D. J. N. J. (2024). Anytime Sequential Halving in Monte-Carlo Tree Search. In M. Hartisch, C.-H. Hsueh, & J. Schaeffer (Eds.), *Computers and Games - 12th International Conference, CG 2024, Virtual Event, November 25-29, 2024, Revised Selected Papers: Computers and Games - 12th International Conference, CG 2024, Virtual Event, November 25-29, 2024, Revised Selected Papers*. https://doi.org/10.1007/978-3-031-86585-5_8

Sala, G., & Gobet, F. (2017). Does Far Transfer Exist? Negative Evidence From Chess, Music, and Working Memory Training. *Current Directions in Psychological Science*, *26*(6), 515–520. https://doi.org/10.1177/0963721417712760

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus Satisficing: Happiness Is a Matter of Choice. *Journal of Personality and Social Psychology*, *83*(5), 1178.

Schwarz, G., Christensen, T., & Zhu, X. (2022,). *Bounded Rationality, Satisficing, Artificial Intelligence, and Decision-Making in Public Organizations: The Contributions of Herbert Simon* (Vol. 82, p. 902). HeinOnline.

Searle, J. R. (1986). *Minds, Brains and Science*. Harvard university press.

Sicart, M. (2022). Thinking the Things We Play With. *Material Game Studies: A Philosophy of Analogue Play*, 21.

Silver, Hubert, et al. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *Corr*, *abs/1712.01815*.

Silver, Schrittwieser, et al. (2017). Mastering the Game of Go without Human Knowledge. *Nat.*, *550*(7676), 354–359. https://doi.org/10.1038/NATURE24270

Sironi, C. F., & Winands, M. H. M. (2016). Comparison of Rapid Action Value Estimation Variants for General Game Playing. *IEEE Conference on Computational Intelligence and Games, CIG 2016, Santorini, Greece, September 20-23, 2016*, 1–8. https://doi.org/10.1109/CIG.2016.7860429

Soemers, D. J. N. J., Mella, V., Piette, É., Stephenson, M., Browne, C., & Teytaud, O. (2023). Towards a General Transfer Approach for Policy-Value Networks. *Trans. Mach. Learn. Res.*, *2023*.

Soemers, D. J. N. J., Piette, É., Stephenson, M., & Browne, C. (2022). Spatial State-Action Features for General Games. *Corr*, *abs/2201.06401*. https://arxiv.org/abs/2201.06401

Sogliuzzo, L., Rautureau, A., & Piette, E. (2026, May). *Toward Modeling Player-Specific Chess Behaviors*.

Stenros, J., & Montola, M. (2024). *The Rule Book: The Building Blocks of Games*. MIT Press.

Stephenson, M., Soemers, D. J. N. J., Piette, É., & Browne, C. (2021). General Game Heuristic Prediction Based on Ludeme Descriptions. *2021 IEEE Conference on Games (CoG), Copenhagen, Denmark, August 17-20, 2021*, 1–4. https://doi.org/10.1109/COG52621.2021.9619052

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning - an Introduction, 2nd Edition*. MIT Press.

Tang, Z., Jiao, D., McIlroy-Young, R., Kleinberg, J. M., Sen, S., & Anderson, A. (2024). Maia-2: A Unified Model for Human-AI Alignment in Chess. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Thagard, P. (1992). Adversarial Problem Solving: Modeling an Opponent Using Explanatory Coherence. *Cognitive Science*, *16*(1), 123–149. https://doi.org/10.1016/0364-0213(92)90019-Q

Thorndike, E. L., & Woodworth, R. S. (1901b). The Influence of Improvement in One Mental Function upon the Efficiency of Other Functions: III. Functions Involving Attention, Observation and Discrimination. *Psychological Review*, *8*(6), 553.

Thorndike, E. L., & Woodworth, R. S. (1901a). The Influence of Improvement in One Mental Function upon the Efficiency of Other Functions. II. The Estimation of Magnitudes. *Psychological Review*, *8*(4), 384.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/MIND/LIX.236.433

Turnwald, A., & Wollherr, D. (2019). Human-Like Motion Planning Based on Game Theoretic Decision Making. *Int. J. Soc. Robotics*, *11*(1), 151–170. https://doi.org/10.1007/S12369-018-0487-2

Tversky, A., Kahneman, D., & Slovic, P. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge.

Vaserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, *5*(3), 64–72.

Vickers, D. (1979,). *Decision Processes in Visual Perception*. Academic Press.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., … Silver, D. (2019). Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nat.*, *575*(7782), 350–354. https://doi.org/10.1038/S41586-019-1724-Z

Von Neumann, J. (1959). On the Theory of Games of Strategy. *Contributions to the Theory of Games*, *4*(13–42), 41.

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in Outlier Detection Techniques: A Survey. *IEEE Access*, *7*, 107964–108000. https://doi.org/10.1109/ACCESS.2019.2932769

Wolff, A. S., Mitchell, D. H., & Frey, P. W. (1984). Perceptual Skill in the Game of Othello. *The Journal of Psychology*, *118*(1), 7–16. https://doi.org/10.1080/00223980.1984.9712586

Woodworth, R. S., & Thorndike, E. L. (1901). The Influence of Improvement in One Mental Function upon the Efficiency of Other Functions.(I). *Psychological Review*, *8*(3), 247.

Yannakakis, G. N., & Togelius, J. (2025). *Artificial Intelligence and Games*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-83347-2

Zhang, Y., Jacob, A. P., Lai, V., Fried, D., & Ippolito, D. (2024, October). *Human-Aligned Chess with a Bit of Search* (Issue arXiv:2410.03893). arXiv. https://doi.org/10.48550/arXiv.2410.03893