university of groningen / faculty of science and engineering / artificial intelligence

CABO - a Computational Analysis of Belief and Opponent-modeling

Bachelor's Project Thesis

Dorin-Vlad Udrea, s4645014, d.udrea@student.rug.nl,
Supervisor: Dr. H.A. de Weerd

Abstract: The human capacity to reason about the unobservable mental states of others, known as Theory of Mind (ToM), is fundamental to navigating complex social interactions. This thesis investigates the strategic value of ToM in a noisy, incomplete-information environment by designing and evaluating a hierarchy of computational agents for the card game CABO. We created a series of AI players with increasingly advanced strategic thinking, beginning with a simple agent that only reacted to observable game events and progressing to complex agents that could model their opponents' hidden beliefs and even anticipate deceptive plays. Agent performance was evaluated in both direct one-on-one competition and complex multi-agent free-for-all games. The results demonstrate a clear, context-dependent performance hierarchy. In pairwise simulations, higher-order ToM confers a significant strategic advantage, with more sophisticated agents consistently outperforming simpler ones. However, this advantage diminishes and even inverts in multi-agent scenarios, where the increased social complexity and cognitive load appear to favor more robust, less complex strategies. The findings confirm that the strategic value of ToM is not absolute but is contingent on the social context and the cognitive capabilities of one's opponents, providing a computational framework for exploring the nuanced dynamics of social intelligence.

1 Introduction

More than two millennia ago, Aristotle famously described humankind as a *zoon politikon*—a social or political animal (Jowett et al., 2000). This observation remains a cornerstone of understanding human nature; our success as a species is deeply rooted in our ability to navigate a complex social world, to cooperate, compete, communicate, and form intricate relationships (Puga-Gonzalez et al., 2009). This social dexterity appears so fundamental that humans will spontaneously attribute intentions, goals, and even personality traits to simple geometric shapes based solely on their movement patterns (Heider & Simmel, 1944). Central to this sociality is the ability to reason about the unobservable mental states of others—their beliefs, desires, and intentions—a capacity commonly referred to as "Theory of Mind" (ToM) (Baron-Cohen et al., 1985). Possessing a Theory of Mind allows us to interpret the actions of others not as mere physical movements, but as the result of internal, psychological motivations, enabling us to predict future behavior, provide altruistic help, and engage in effective social interaction (Warneken & Tomasello, 2006).

A crucial benchmark for assessing ToM is the false-belief task, designed to test the understanding that others can hold beliefs that conflict with reality (Baron-Cohen et al., 1985). False-belief tasks are considered a gold standard in ToM research because they require a clear distinction between one's own knowledge and the mental representation of another person (Kosinski, 2024; Wellman et al., 2001). In the classic "Sally-Anne" task, for example, a child must predict that an agent will act upon her outdated, false belief about an object's location, rather than its true location (Baron-Cohen et al., 1985).

Research has consistently documented a clear developmental trajectory for this ability. A comprehensive meta-analysis of 178 studies confirmed that while children younger than three and a half years old systematically fail these tasks, there is a fundamental conceptual shift between the ages of three and five, after which most children consistently pass (Wellman et al., 2001). This developmental pattern is remarkably robust across different cultures and languages (Wellman et al., 2001). However, this ability is not monolithic. Studies have shown a developmental lag, where the ability to attribute simple ignorance (that someone does not know a fact) precedes the more complex ability to attribute a specific false belief (Hogrefe et al., 1986). This suggests that representing conflicting states of affairs (the world as it is vs. the world as it is believed to be) is a key conceptual hurdle. Furthermore, various task factors can influence performance rates; for instance, framing the task in terms of active deception or requiring the child's direct participation can improve success rates, though these factors do not typically eliminate the underlying developmental trend (Wellman et al., 2001).

This developmental progression should not imply that adult ToM is flawless. Even for adults, reasoning about the beliefs of others can be compromised by their own knowledge. This cognitive bias, known as the "curse of knowledge," can lead adults to overestimate what a less-informed person knows, especially when a plausible (but incorrect) rationale for the other's action is available (Birch & Bloom, 2007). Furthermore, the way individuals use ToM is connected to how they explain social events; deficits in mentalizing ability have been linked to an increased tendency to make external-personal causal attributions, blaming other people for negative outcomes rather than situational factors (Kinderman et al., 1998). These nuances have led to critical examinations of the experimental paradigms themselves. For example, some have questioned whether the widely-used "director task" is a pure test of ToM. This task is a type of referential communication game where a participant, the "follower", is instructed by a "director" to move objects within a vertical grid of squares. The crucial element is that the two individuals have different perspectives; some of the squares are occluded from the director's view, so she cannot see all the objects that the follower can. The follower is aware of the director's limited knowledge. In a classic example, a director might ask the follower to "move the small candle" when there are three candles of different sizes, but the smallest of the three is visible only to the follower. A follower using Theory of Mind would correctly infer that the director must be referring to the medium-sized candle, as it is the smallest one from her perspective. However, studies often show that participants exhibit an "egocentric bias," where they consider or even reach for the object in their own privileged view before selecting the correct one. The central critique is that optimal performance could be achieved through domain-general selective attention alone, without necessarily representing the director's mental state, as a participant could simply adopt a strategy to ignore the occluded cells altogether (Rubio-Fernández, 2017).

The question of whether ToM is a uniquely human ability has fueled extensive research in comparative psychology. Studies on our closest primate relatives, chimpanzees, reveal a complex picture. Chimpanzees possess some foundational ToM skills; they can understand what conspecifics can and cannot see and use this knowledge in competitive situations (Hare et al., 2001). They may even demonstrate altruistic helping behaviors based on an understanding of others' goals (Warneken & Tomasello, 2006). However, despite these abilities, chimpanzees have consistently failed more complex false-belief tests that school-aged children pass, even when the tests are designed to be more species-relevant and non-verbal (Krachun et al., 2010). This debate extends to other socially intelligent species. For instance, complex behaviors like food re-caching in corvids can be interpreted as either evidence of ToM or, more parsimoniously, as the result of simpler associative learning mechanisms and stress responses (Van Der Vaart et al., 2012). Similarly, complex primate affiliation patterns, such as grooming and reconciliation in chimpanzees, can emerge from simple, local rules based on anxiety and proximity, without requiring sophisticated social cognition (Puga-Gonzalez et al., 2009).

More recently, the investigation of ToM has extended into the domain of artificial intelligence, particularly with the advent of Large Language Models (LLMs). This emerging field of "machine psychology" seeks to evaluate whether these models can replicate human-like social reasoning by subjecting them to a battery of psychological tests (Strachan et al., 2024). The results are mixed but tantalizing. Some studies suggest that ToM-like abilities may have spontaneously emerged in LLMs, showing that models like GPT-4 can solve a high percentage of false-belief tasks at a level comparable to young children (Kosinski, 2024). Indeed, some LLMs can even outperform humans in specific social situational judgment tests, correctly identifying socially appropriate behaviors in complex scenarios (Mittelstädt et al., 2024). However, this capability appears brittle. Other work has demonstrated that LLMs often fail when presented with trivial alterations to the standard task formats, such as making a container transparent, suggesting their success may rely on shallow heuristics rather than genuine mentalistic inference (Ullman, 2023). The failure of some models, such as GPT's difficulty with "faux pas" detection, has been attributed not to a failure of inference but to a "hyperconservative" response strategy, where the model avoids committing to a conclusion in the face of uncertainty (Strachan et al., 2024). This contrasts with human-robot interaction studies, where a robot that clearly "cheats" in a game is more likely to have mental states like intentionality attributed to it by human participants than a robot that simply malfunctions (Short et al., 2010).

To better understand the mechanisms and evolutionary advantages of ToM, researchers have utilized agent-based computational models. This approach often formalizes ToM as a problem of inverse planning, where an agent infers the hidden beliefs and desires of another by assuming they are a rational actor trying to achieve a goal (Devaine et al., 2014a). This leads to so-called "meta-Bayesian" agents that reason about other agents who are, themselves, reasoning beings (Devaine et al., 2014b, 2017). This framework allows for precise control over an agent's reasoning abilities, providing a testbed for exploring the impact of ToM on social outcomes. For instance, simulations have shown that higher-order ToM—the ability to reason about what others believe about your beliefs—confers a significant advantage in competitive games (De Weerd et al., 2013) and is essential for navigating mixed-motive scenarios with incomplete information (De Weerd et al., 2017). However, other work adds crucial nuance, demonstrating that the evolutionary fitness of ToM is highly context-dependent. While beneficial in competition, cooperative settings can surprisingly favor lower levels of ToM sophistication, suggesting that there is no absolute evolutionary pressure for ever-increasing mentalizing complexity (Devaine et al., 2014b). Indeed, these computational approaches have been used to show that a species' general cognitive capacity, rather than its social complexity, may be the primary driver of its ToM evolution (Devaine et al., 2017). This body of work strongly supports the hypothesis that our sophisticated social cognition evolved to navigate complex social dynamics, but highlights that the optimal level of reasoning is a fine-tuned balance of cognitive capacity and social context.

The limitations of static, puzzle-like scenarios for evaluating advanced ToM in adults and LLMs suggest a need for new, more dynamic paradigms. Strategic games of incomplete information, which require continuous belief updating and opponent modeling, offer a promising alternative. The card game CABO, in particular, presents an ideal environment for this purpose. Its rules—involving peeking at one's own cards, spying on others, swapping cards, and strategically calling "CABO" based on an assessment of all players' hands—create a rich context for the strategic use of Theory of Mind. Success in CABO is not merely about optimizing one's own hand; it requires actively modeling the knowledge and beliefs of other players.

However, before CABO can be established as a robust paradigm for testing ToM in humans or AI, a foundational question must be answered: is ToM a demonstrably useful strategy within the formal structure of the game itself? This thesis addresses this primary question by computationally formulating how ToM can be applied in CABO. By designing and implementing a hierarchy of artificial agents with varying levels of ToM, this research investigates whether modeling the beliefs and knowledge of opponents provides a measurable advantage in this game of incomplete information and social deduction. This work therefore serves as a necessary first step: a computational analysis to validate the strategic relevance of opponent modeling in CABO.

The remainder of this thesis is structured as follows. Chapter 2 introduces the rules of the card game CABO and formalizes its strategic elements. It also describes the architecture of the ToM agent, detailing the mechanisms for belief tracking and decision-making. Chapter 3 presents the results of simulations comparing the performance of the ToM agents against each other. Finally, Chapters 4 and 5 discuss the implications of these findings, address the limitations of the current work, and outline promising directions for future research.

2 Methods

To investigate the role and efficacy of different orders of ToM in a strategic environment, a computational model was developed centered on the card game CABO. This section details the theoretical framework underpinning the model, the architecture of the cognitive agents, and the methods used for their evaluation. The approach is grounded in established cognitive science principles of ToM, implementing a hierarchy of agents capable of increasingly sophisticated social reasoning.

2.1 The Game Environment: CABO

The card game CABO was selected as the experimental testbed for this research due to its properties as a mixed-motive game of incomplete information. In this game, which accommodates two to four players, the objective is purely competitive. Players aim to achieve the lowest score by managing a hand of four face-down cards from a standard deck of playing cards, and only the single player with the lowest total value wins the round. Gameplay proceeds in a clockwise order. On their turn, a player must perform one of three actions:

- Draw a card from the deck. The player looks at the card they have drawn and must choose one of the following two options:
  - Swap the drawn card with one of their own face-down cards. The player chooses which of their four cards to replace, typically without knowing its value unless they have used a "peek" action previously. The replaced card goes into the discard pile.
  - Discard the drawn card onto the face-up discard pile.
- Take the top card from the face-up discard pile and swap it with one of their own face-down cards.
- Call "CABO" to initiate the end of the round.

A player may discard multiple cards from their hand simultaneously if they are of the same rank; if this results in an empty hand, their score is fixed at zero. When a player calls "CABO," they do not alter their hand, and all other players get one final turn. During these final turns, players may use any available actions, including "swap," which can target any card on the table, including those in the hand of the player who called "CABO." If the draw deck runs out of cards, the current player is forced to call "CABO," and the game proceeds to the final round of turns. To prevent games from running indefinitely, which occurred in rare stalemates when agents performed actions that did not deplete the draw deck (e.g., using powers), a 100-turn limit is enforced in simulations; if this limit is reached, the game is declared a tie.

If any card a player moves to the discard pile during their turn—either by discarding a freshly drawn card or by replacing a card in their hand—has a special value, that player may use the corresponding ability on a future turn. The abilities are: cards 7 or 8 allow a player to "peek" at one of their own cards; 9 or 10 allow a "spy" on an opponent's card; and 11 or 12 allow a "swap" of any two cards on the table.

Key features of CABO make it an ideal environment for studying ToM:

- Players have only partial knowledge of the game state. Each player begins by peeking at only two of their own four cards, and the contents of the draw pile and opponents' hands are unknown. This uncertainty necessitates that players form beliefs about the hidden aspects of the world.
- The game's actions directly involve reasoning about knowledge. Actions such as "Spy" or "Swap" are explicitly designed to manipulate the knowledge states of the participants.
- The state of knowledge is constantly in flux. Every action provides information that can be used to update one's beliefs about the game state.

This environment forces agents beyond simple, reactive strategies and compels them to engage in the kind of belief- and goal-based reasoning that is the hallmark of Theory of Mind.

2.2 A Hierarchical Model of Theory of Mind

The primary method involves the implementation of a hierarchy of computational agents, each corresponding to a different level of ToM reasoning, from zero-order (non-mentalistic) to third-order (ToM_0 – ToM_3). This hierarchical approach allows for a controlled comparison of the strategic advantages conferred by deeper social cognition. The conceptual logic of each agent is described below, with the formal decision-making model detailed in Section 2.3.

2.2.1 The Zero-Order Agent (ToM_0)

The ToM_0 or "zero-order" agent evaluates actions based on their immediate, selfish outcomes. To handle incomplete information, it generates a belief distribution over possible game states ("worlds") consistent with its own knowledge. It then calculates the expected value of each action by simulating them across these worlds. Its reasoning is zero-order because the simulation stops after its own move; it calculates its score immediately and does not model any future actions by its opponents. For instance, it will take a known low-value card from the discard pile if that action has a better expected outcome for its own score than drawing an unknown card, a decision that does not involve reasoning about opponents' mental states.

2.2.2 The First-Order Agent (ToM_1)

The ToM_1 agent represents the first level of recursive mentalistic reasoning, operating on a fixed architecture: it always models its opponent as a selfish, baseline ToM_0 agent. To select an action, it simulates the game one step further to forecast its opponent's probable response. This foresight allows it to avoid costly errors that a simpler agent would make.

For example, if a ToM_1 agent holds a very low-value card (e.g., a 1), it anticipates that drawing a new card might force it to discard the 1. Its simulation predicts that a ToM_0 opponent would then certainly take the discarded card. By modeling the opponent's selfish behavior, the ToM_1 agent foresees this "double penalty"—worsening its own score while improving its opponent's—and can thus select a more effective counter-action.

However, this fixed model becomes a significant weakness in complex games. In multi-agent, free-for-all scenarios with a mix of agent types, the assumption that every opponent is a ToM_0 is often incorrect, a factor that negatively impacts the ToM_1 agent's performance.

2.2.3 The Second-Order Agent (ToM_2)

The ToM_2 agent takes the recursion one step further, modeling its opponent as a ToM_1 agent. This allows it to reason not just about the opponent's beliefs about the world, but about the opponent's beliefs *about its own beliefs and intentions.* This capability enables more complex strategic behaviors like deception.

For instance, a ToM_2 agent might take a known high-value card (e.g., a 12, which grants a "swap" power) from the discard pile. A simpler agent would avoid this, as it worsens its score. However, the ToM_2 agent's simulation includes its opponent's reasoning: it *believes* that the ToM_1 opponent will *think* the card was taken with the *intention* of using its powerful swap ability on a future turn. By manipulating the opponent's beliefs about its future plans, the ToM_2 agent may cause the opponent to delay calling "CABO," giving it more turns to genuinely improve its hand.

2.2.4 The Third-Order Agent (ToM_3) and Model Limitations

The highest level of reasoning implemented is the ToM_3 agent, which models its opponent as a ToM_2 agent. The choice was made to cap the hierarchy at the third order for two primary reasons. First, computational complexity grows exponentially with agent level. Second, empirical research on human cognition suggests that the strategic benefits of ToM tend to plateau beyond the second order in strategic tasks (De Weerd et al., 2013).

2.3 Formal Model of Agent Decision Making

To act under uncertainty, all ToM agents employ a belief state simulation to approximate the true state of the game. On its turn, an agent generates a belief distribution, $\mathcal{B}$, consisting of a set of N_{worlds} plausible game states, or "worlds."

These worlds are sampled to be consistent with the agent's current knowledge. The sampling process begins with a full 52-card deck. First, all cards known to the agent—including its own peeked cards and all cards in the public discard pile—are removed from this deck to create a pool of all possible unknown cards. A single "world" is then generated by randomly dealing cards from this pool into all unknown slots (the agent's un-peeked cards and all opponents' hands). The remaining cards in the pool form the draw pile for that world.

This process is repeated to generate N worlds, where $N_{worlds} = 1000$. This number was chosen as a balance between computational performance and the accuracy of the agent's world model.

The decision of which action to take is then made by calculating the expected utility, $Q(a)$, for each legal action a over this belief space. Since the goal in CABO is to minimize one's score, utility is defined as the negative of the expected final score. The expected utility is given by:

$$Q(a) = \sum_{w \in \mathcal{B}} P(w) \cdot U(s'_w) \tag{2.1}$$

where $P(w)$ is the probability of a given world w (assumed to be uniform across the sampled worlds), and $U(s'_w)$ is the utility of the game state resulting from taking action a in world w. This resulting utility is determined through a recursive simulation, as detailed in Algorithm A.1 in the Appendix.

2.4 Experimental Setup and Evaluation

To evaluate agent performance, agents were simulated in two settings: pairwise (1-vs-1) games and multi-agent (4-player, free-for-all) games. The key parameters used throughout all simulations are detailed in Table 2.1.

2.4.1 Procedure and Metrics

In each simulated game, the turn order of agents was randomly shuffled to prevent first-turn advantage. The primary performance metric analyzed was the Win Rate, defined as the proportion of games won by an agent. Additional metrics, including average scores and game length, are detailed in the Appendix.

Table 2.1: Model and Simulation Parameters

Parameter	Value	Description
Games per Experiment	1000	Number of games simulated for each pairing/group.
Belief Worlds (N_{worlds})	1000	Number of worlds sampled per decision.
Epsilon (ϵ)	0.05	Probability of choosing a random action (exploration).
Beta (β)	1.0	Inverse temperature for softmax decision-making.

2.4.2 Analysis and Statistical Testing

The collected data were analyzed using several methods. To compare the explanatory power of the different agent models, Random-Effects Bayesian Model Selection (RFX-BMS) was employed, complemented by the Akaike (AIC) and Bayesian (BIC) Information Criteria.

To determine if observed differences in win rates were statistically meaningful, Chi-squared (χ^2) tests were used. For pairwise contests, a goodness-of-fit test was used to see if win rates deviated significantly from chance (a 50/50 split of non-tied games). For multi-agent games, a test of independence was used to determine if there was a significant association between agent type and the frequency of winning. These tests were chosen as they are appropriate for analyzing categorical outcome data (win/loss). A significance level of $p < 0.05$ was used for all tests.

3 Results

This chapter presents the empirical results obtained from simulating the implemented agents in the game of CABO. The performance of agents with varying orders of Theory of Mind (ToM_1, ToM_2, ToM_3) was compared against each other and against a baseline ToM_0 agent. The primary performance metric analyzed in this section is the win rate of each agent. Additional metrics, such as average scores and game length, are detailed in the Appendix. Performance was evaluated in two distinct settings: pairwise (1-vs-1) games and multi-agent (free-for-all) games.

3.1 Pairwise Agent Performance

To assess the strategic advantage of ToM in a direct competitive context, agents were paired against each other in one-on-one games. The mean win rate for each agent type against every opponent type was calculated from the simulation data. The results are summarized in Table 3.1.

Table 3.1: Mean win rates from pairwise (1-vs-1) agent simulations. Each cell represents the win rate of the row agent against the column agent. Bold values with an asterisk (*) indicate a statistically significant advantage ($p < 0.05$).

Agent	ToM_3	ToM_2	ToM_1	ToM_0
ToM_3	-	0.517*	0.463	0.505
ToM_2	0.437*	-	0.482	0.502
ToM_1	0.493	0.461	-	0.496
ToM_0	0.456	0.446	0.447	-

The results in Table 3.1 generally reveal a clear hierarchical relationship based on the order of ToM, where higher-order agents demonstrate an advantage over lower-order ones. This trend can be seen by reading across the rows; for instance, ToM_2 achieved a win rate of 0.502 against ToM_0, and ToM_3 beat ToM_2 with a win rate of 0.517. A series of Chi-squared tests, detailed in the Appendix, confirms that the advantage of ToM_3 over ToM_2 is statistically significant ($\chi^2(1) = 6.71, p < 0.01$). It should be noted that win rates do not sum to 1.000 due to games ending in a tie. A tie was declared if a game reached the 100-round simulation limit without any player calling “CABO.” This occurred in rare stalemates where agents performed actions that did not deplete the draw deck (e.g., using powers), leading to a state where no agent gained sufficient confidence to end the round.

However, there is a notable exception to this hierarchy: the most sophisticated agent, ToM_3, was less effective against the simpler, information-gathering ToM_1 agent, achieving a win rate of only 0.463. This suggests a potential cognitive cost to “overthinking,” where a complex model designed to counter high-level deception is mismatched and ineffective against a more direct strategy.

3.2 Multi-Agent Game Performance

To evaluate performance in a more complex social environment, agents were simulated in multi-agent (free-for-all) games. Table 3.2 shows the mean win rate for each agent type in this setting.

Table 3.2: Mean win rates in multi-agent (free-for-all) games.

Agent	WinRate
ToM_3	0.249
ToM_2	0.223
ToM_1	0.234
ToM_0	0.259

In stark contrast to the pairwise results, the performance hierarchy disappears in the multi-agent games. This is likely because the agents are implemented with a fixed, rigid model of their opponents (e.g., a ToM_2 agent always assumes it is playing against a ToM_1 agent). In a complex, mixed-agent environment, this model is frequently incorrect, increasing the cognitive load and favoring less complex, more robust strategies. The data in Table 3.2 shows this trend is even inverted, with the baseline ToM_0 agent achieving the highest win rate (0.259). Notably, the performance is not linear, as the ToM_2 agent records the lowest win rate. However, a Chi-squared test of independence found that these differences in win rates are not statistically significant ($\chi^2(3) = 4.16, p = 0.25$), suggesting that no single agent type had a discernible advantage in the multi-agent setting.

4 Discussion

This thesis set out to investigate the strategic value of Theory of Mind (ToM) in the complex, incomplete-information card game of CABO. By implementing a series of computational agents with varying orders of ToM and simulating their performance, this study provides empirical evidence for the role of mental state attribution in strategic play. This penultimate chapter interprets the findings presented in the Results section, discusses their implications in the context of the broader literature, acknowledges the limitations of the current work, and proposes directions for future research.

4.1 Interpretation of Agent Performance

The simulation results demonstrate a clear, albeit nuanced, performance hierarchy based on an agent's ToM capabilities. The advantage of modeling opponents' minds is most apparent in direct competition, but the dynamics shift significantly in more complex multi-agent scenarios, highlighting the context-dependent nature of cognitive strategies.

4.1.1 The Strategic Hierarchy and the Cost of Overthinking in Pairwise Games

In one-on-one games, the results provide the clearest evidence for the value of recursive mental modeling. The performance of each agent type can be interpreted as an escalating series of strategies and counter-strategies that emerge directly from their underlying simulation-based reasoning, as described in the Methods section:

- The ToM_0 agent serves as a baseline. Its core heuristic is a myopic, utility-maximizing calculation based on its own immediate score. Its general failure against other agents underscores the inadequacy of a strategy that does not anticipate an opponent's response.

- The ToM_1 agent moves beyond this by "playing the player." Its advantage comes from its one-level recursion, which allows it to anticipate the probable counter-move from a selfish ToM_0 opponent and thereby make more cautious and better-informed decisions.

- The ToM_2 agent leverages this for deception. It "plays the player's game" by exploiting the ToM_1 agent's reasoning process. Its two-level recursion allows it to identify actions that can mislead a ToM_1 opponent, for example about its own hand strength, to gain a strategic advantage.

- The ToM_3 agent anticipates this deception. By modeling its opponent as a ToM_2 agent, its three-level recursion can better distinguish actions that are genuinely beneficial from those that are likely deceptive bluffs, allowing it to "call the bluff" more effectively.

This theoretical hierarchy is largely supported by the pairwise results in Table 3.1, where higher-order agents generally outperform lower-order ones. However, the most surprising and arguably most insightful result is the performance of ToM_3 against ToM_1. The ToM_1 agent achieves a win rate of nearly 50% (0.493), effectively neutralizing the more complex ToM_3 agent. This suggests that ToM_3's sophisticated model, which is calibrated to anticipate the bluffs and deceptions of a ToM_2 agent, represents a form of cognitive "overthinking" when faced with the more direct, information-gathering strategy of a ToM_1 opponent. The ToM_1 agent does not engage in the kind of elaborate deception that ToM_3 is designed to counter, making the ToM_3's complex model a poor and inefficient fit for its opponent's simpler strategy.

4.1.2 Performance Saturation in Multi-Agent Environments

In free-for-all games, the clear hierarchical advantage observed in pairwise contests disappears, with win rates for all agent types clustering around the chance level of 25% (Table 3.2). This suggests a performance saturation effect in complex social environments. The strategic advantage a ToM_n agent holds over a ToM_{n-1} opponent is often nullified by the presence of other agents. As the ToM agents are implemented with a fixed, static model of their opponents, this model is frequently mismatched in a heterogeneous environment (e.g., a ToM_2 agent assumes it is playing against ToM_1 agents, but is actually facing a mix of ToM_0, ToM_1, and ToM_3 agents). The cognitive resources required to run simulations against a frequently incorrect model leads to a noisier and less reliable predictive process. In this chaotic environment, the specific advantage of out-predicting one opponent is diminished, as the actions of other players introduce significant unpredictability. The strategic landscape becomes a dynamic equilibrium where the benefits of higher-order ToM are largely canceled out, leading to the observed plateau in performance. While win rates equalize, it is worth noting that average scores (detailed in the Appendix) still show a slight advantage for higher-ToM agents, suggesting they are more effective at damage control and minimizing losses, even if they do not secure more outright wins.

4.2 Limitations and Future Work

While this study successfully demonstrates the context-dependent value of ToM in CABO, it is important to acknowledge its limitations, which in turn suggest avenues for future research.

4.2.1 Scope of the Agent Model and Computational Complexity

The ToM agent developed for this thesis, while effective, is a simplified model of social reasoning. One key limitation is its fixed, static model of opponents. A promising direction for future work would be to create a more flexible agent capable of inferring the reasoning level of its specific opponents and adapting its strategy accordingly.

Furthermore, the "performance saturation" observed in multi-agent games warrants deeper investigation. A valuable next step would be to conduct simulations in homogenous multi-agent environments to disentangle the effects of cognitive load from model mismatch. For instance, testing a single ToM_3 agent against a table of three ToM_2 agents would create a scenario where its internal model is perfectly correct. We would hypothesize that in such an environment, the strategic advantage of higher-order ToM would reappear, providing stronger evidence for the importance of accurate opponent modeling.

Finally, several advanced computational concepts were not implemented but represent promising directions for future work:

- The model's decision-making relies on a depth-limited expectimax search performed across the sampled belief space. For each possible action, it recursively calculates the opponent's optimal counter-move up to a depth determined by its ToM level (e.g., a ToM_2 agent simulates two opponent moves). This approach is limited in two key ways: the search is exhaustive at each step, making it computationally inefficient for exploring deep game sequences, and the fixed search depth creates a "horizon effect" where the agent cannot foresee consequences beyond its immediate simulation. More advanced search algorithms like Monte Carlo Tree Search (MCTS) could overcome these limitations by intelligently sampling promising game pathways to explore the game tree more efficiently and deeply. Furthermore, implementing Reinforcement Learning (RL) would allow an agent to learn an optimal policy from experience, moving beyond the current static utility calculations to discover novel, non-intuitive strategies.

- Decision-making under uncertainty is central to CABO. While belief-space sampling is a pragmatic approach, a more formal and robust framework could be achieved using Partially Observable Markov Decision Processes (POMDPs). These models are explicitly designed for such problems, though they are notoriously complex to solve and would require significant computational resources.

- The current model updates its beliefs through resampling. A more structured approach could use Bayesian Belief Networks to represent and update the probabilistic beliefs about opponents' hands and intentions in a more principled and potentially more accurate manner, explicitly modeling the conditional dependencies between different pieces of information.

- The current agents model opponents' beliefs and goals with the singular motive of winning (i.e., minimizing their score). The model does not account for other potential motives, such as risk-aversion, spite, or a preference for cooperation, which are relevant in mixed-motive games. Expanding the model to include these factors could lead to more nuanced and human-like agent behavior.

Integrating these techniques represents a significant and promising direction for future research. Overcoming the computational challenges they present could lead to the development of a more adaptive, robust, and ultimately more formidable CABO agent, providing deeper insights into the nature of strategic social cognition.

5 Conclusion

This thesis successfully designed, implemented, and evaluated a series of computational agents for the game of CABO, confirming that Theory of Mind is a powerful, yet context-dependent, mechanism for navigating the game's complex social and strategic landscape. The core contribution of this work is the empirical demonstration of a cognitive hierarchy in a noisy, incomplete-information game environment. A clear performance advantage was established for agents capable of recursively modeling the knowledge and beliefs of their opponents, particularly in direct one-on-one competition. This finding aligns with and provides further support for existing research in computational modeling, which has consistently shown that higher-order ToM confers a significant strategic benefit in both competitive and mixed-motive scenarios (De Weerd et al., 2013, 2017).

Furthermore, this study revealed a nuanced and surprising dynamic: the effectiveness of a particular ToM level is contingent on the cognitive sophistication of its opponent. The near-50% win rate of the simpler ToM_1 agent against the more complex ToM_3 agent suggests that very high orders of reasoning can become a form of "overthinking"—a cognitive mismatch where a model designed to anticipate complex deception is ineffective against a more direct, information-gathering strategy. This highlights a critical principle in social intelligence: optimal strategy is not about maximizing cognitive complexity in isolation, but about appropriately calibrating one's own model to the mind of the specific other.

In more complex multi-agent settings, the clear advantage of higher-order ToM seen in pairwise games was attenuated, with performance across all agent types converging. This suggests a "performance saturation" effect, where the cognitive load of modeling multiple, heterogeneous minds and the increased unpredictability of the environment diminish the strategic returns of deep recursive reasoning. In such chaotic social ecologies, simpler and more robust heuristics can prove surprisingly effective.

This work contributes a practical application of ToM principles to a real-world recreational game, providing a robust, open-source framework for future research into artificial social intelligence. The limitations of the current model—its reliance on simplified learning mechanisms and the absence of more advanced AI architectures like Reinforcement Learning or POMDPs—do not detract from its findings but rather illuminate a clear path forward. The developed CABO environment can serve as a valuable and challenging testbed for exploring these more advanced learning algorithms and for further investigating the intricate dynamics of machine Theory of Mind. By continuing to bridge the gap between abstract models and the complexities of real-world social interaction, this line of research promises to deepen society's understanding of the computational underpinnings of social intelligence, both in humans and in the increasingly sophisticated artificial agents that are created.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386.

Devaine, M., Hollard, G., & Daunizeau, J. (2014a). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, *10*(12), e1003992.

Devaine, M., Hollard, G., & Daunizeau, J. (2014b). Theory of mind: Did evolution fool us? *PloS One*, *9*(2), e87619.

Devaine, M., San-Galli, A., Trapanese, C., Bardino, G., Hano, C., Saint Jalme, M., . . . Daunizeau, J. (2017). Reading wild minds: A computational assay of theory of mind sophistication across seven primate species. *PLoS Computational Biology*, *13*(11), e1005833.

De Weerd, H., Verbrugge, R., & Verheij, B. (2013). How much does it help to know what she knows you know? an agent-based simulation study. *Artificial Intelligence*, *199*, 67–92.

De Weerd, H., Verbrugge, R., & Verheij, B. (2017). Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, *31*, 250–287.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 567–582.

Jowett, B., Davis, H. W. C., et al. (2000). *Politics*. Courier Corporation.

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121.

Krachun, C., Call, J., Tomasello, M., et al. (2010). A new change-of-contents false belief test: Children and chimpanzees compared. *International Journal of Comparative Psychology*, *23*(2).

Mittelstädt, J. M., Maier, J., Goerke, P., Zinn, F., & Hermes, M. (2024). Large language models can outperform humans in social situational judgments. *Scientific Reports*, *14*(1), 27449.

Puga-Gonzalez, I., Hildenbrandt, H., & Hemelrijk, C. K. (2009). Emergent patterns of social affiliation in primates, a model. *PLoS Computational Biology*, *5*(12), e1000630.

Rubio-Fernández, P. (2017). The director task: A test of theory-of-mind use or selective attention? *Psychonomic Bulletin & Review*, *24*(4), 1121–1128.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 219–226).

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... others (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Van Der Vaart, E., Verbrugge, R., & Hemelrijk, C. K. (2012). Corvid re-caching without 'theory of mind': A model. *PloS one*, *7*(3), e32904.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

A Appendix

Algorithm A.1 Recursive Action Valuation

```
function CalculateValue(action, agent, world, level)
    Simulate action by agent in world to get newWorld
    if IsTerminal(newWorld) then
        return -GetScore(agent, newWorld)
    end if
    nextPlayer := GetNextPlayer(newWorld)
    if level > 0 then
        OpponentModel := InitializeAgent(level-1)
        oppAction := OpponentModel.Decide(newWorld)
    else
        oppAction := SampleFromBehavioralModel(nextPlayer)
    end if
    return CalculateValue(oppAction, nextPlayer, newWorld, level)
end function
```

Table A.1: Free Metrics

Agent	WinRate	AvgScore	ScoreStd	AvgRank	AvgTurns	RFX	AIC	BIC
ToM_0	0.259	21.250	11.793	2.468	19.536	0.25	22 615.975	22 635.606
ToM_1	0.234	21.439	11.258	2.503	19.536	0.25	23 661.564	23 681.195
ToM_2	0.223	21.682	10.955	2.532	19.536	0.25	23 441.403	23 461.034
ToM_3	0.249	21.108	11.716	2.497	19.536	0.25	27 025.142	27 044.773

Table A.2: Pairwise Metrics (Part 1 of 2)

Agent A	Agent B	WinRate A	WinRate B	TieRate	AvgScore A	AvgScore B	ScoreStd A
ToM_0	ToM_1	0.447	0.496	0.057	19.438	18.226	15.623
ToM_0	ToM_2	0.446	0.502	0.052	20.101	18.638	15.628
ToM_0	ToM_3	0.456	0.505	0.039	19.832	18.955	15.119
ToM_1	ToM_2	0.461	0.482	0.057	18.956	17.277	14.539
ToM_1	ToM_3	0.493	0.463	0.044	18.440	18.492	14.272
ToM_2	ToM_3	0.437	0.517	0.046	19.770	17.205	14.565

Table A.3: **Pairwise Metrics (Part 2 of 2)**

Agent A	Agent B	ScoreStd B	AvgTurns	RFX A	RFX B	AIC A	AIC B	BIC A	BIC B
ToM_0	ToM_1	15.288	17.116	0.503	0.497	25 875.035	33 446.751	25 894.667	33 466.382
ToM_0	ToM_2	15.242	15.685	0.501	0.499	23 208.622	32 700.425	23 228.253	32 720.056
ToM_0	ToM_3	15.459	15.429	0.474	0.526	24 920.258	32 711.863	24 939.889	32 731.494
ToM_1	ToM_2	15.516	18.006	0.522	0.478	31 281.723	36 510.406	31 301.354	36 530.037
ToM_1	ToM_3	15.775	18.071	0.490	0.510	34 546.873	37 856.671	34 566.504	37 876.302
ToM_2	ToM_3	15.520	16.611	0.496	0.504	32 783.127	35 189.068	32 802.758	35 208.699

Table A.4: **Full Statistical Test Results for Pairwise Matchups**

Agent A	Agent B	χ^2 Statistic	p-value	Significant (p<0.05)
ToM_0	ToM_1	2.55	0.1106	No
ToM_0	ToM_2	3.31	0.0689	No
ToM_0	ToM_3	2.50	0.1140	No
ToM_1	ToM_2	0.47	0.4941	No
ToM_1	ToM_3	0.94	0.3319	No
ToM_2	ToM_3	6.71	0.0096	Yes