Cognition 266 (2026) 106331

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Strategic reasoning under pressure: Testing heuristics in higher-order theory of mind

Gregory N. Stanley *

University of Michigan, Department of Psychology, Cognition and Cognitive Neuroscience, 530 Church Street, Ann Arbor, MI 48109-1043, USA

ARTICLE INFO

Keywords:
Higher-order theory of mind
behavioral game theory
k-level models
Cognitive load
Heuristics

ABSTRACT

Higher-order Theory of Mind (ToM+)—the recursive ability to understand that others have thoughts about thoughts—is pivotal to complex social interactions but can be cognitively demanding. This study examines how individuals cope when ToM+ reasoning exceeds their cognitive limits, contrasting a model predicting "blindness" to future states with a model predicting averaging over future states. Conducted on a new online platform called the *Morality Game* and using a series of 32 time-pressured sequential-choice games, participants' errors reveal consistent support for the probabilistic model, indicating that when precise higher-order reasoning is unfeasible, individuals do not simply ignore future possibilities. Instead, they approximate future states using probabilistic heuristics rather than explicit recursive reasoning. By clarifying how individuals rely on such heuristics when full deliberation is impractical, these findings provide a clearer framework for understanding the mental shortcuts that shape ToM+ under cognitive strain, thereby informing more precise methods for investigating and modeling complex social inference.

1. Introduction

Theory of Mind (ToM) is the capacity to represent what others think, feel, want, or intend (Estes, Wellman, & Woolley, 1989; Gopnik & Wellman, 1992; Perner, 1994; Premack & Woodruff, 1978; Wellman, Carey, Gleitman, Newport, & Spelke, 1990). By contrast, higher-order Theory of Mind (ToM+) involves representing not just others' thoughts, but others' thoughts about thoughts—recursive mental representations extending multiple levels deep. ToM operates when an individual tries to impress another person, but ToM+ operates when an individual tries to impress another person while hiding this intent. Similarly, while basic ToM suffices for knowing a lover's desire, ToM+ is needed to know if that lover knows if that desire is reciprocated. Such recursive mentalizing undergirds many nuanced social phenomena. Sarcasm hinges on the speaker's awareness that the listener can detect the intended contradiction between words and meaning, while dramatic irony relies on an author's recognition that the audience knows more than a character does. Even large-scale strategic standoffs, such as the Cold War's mutual nuclear deterrence, depended on each side's understanding of the other's understanding—several layers deep—of the catastrophic consequences of a first strike.[1] Thus, ToM+ proves critical to navigating complex social interactions, from intimate relationships and everyday banter to the geopolitical balance of power (McNamara, 1962; Schelling, 1960; Waltz, 1981).

This study leverages the mathematical framework of Game Theory, which represents multi-agent decisions with game trees: nodes represent states, edges represent state transitions, branching nodes represent decisions, and leaf nodes represent outcomes with payoffs. Fig. 1 shows an example sequential-choice game from my study, played from the top (root node) downward. Player 1 (red-blue sun-shaped avatar) first chooses to stay (ending with payoffs 2 for player 1 and 3 for player 2) or move, in which case player 2 (blue octagon avatar) makes a similar choice. If player 2 moves, player 1 again chooses among two outcomes. Players aim to selfishly maximize their own payoffs in each game, controlling for social preferences (e.g., altruism) and cross-game strategic considerations (e.g., reciprocity).

Even this limited form of ToM, where everyone shares identical goals and beliefs, still demands recursive reasoning about other minds: at the

* Corresponding author.
E-mail address: gregstan@umich.edu.

[1] Recent agent-based work shows that higher-order ToM is also especially advantageous in unpredictable negotiations (De Weerd, Verbrugge, & Verheij, Higher-order theory of mind is especially useful in unpredictable negotiations, 2022). Likewise, interlocutors only succeed in some tacit-communication tasks when they can predict how their moves will be interpreted several levels deep (De Weerd, Verbrugge, & Verheij, 2015).

https://doi.org/10.1016/j.cognition.2025.106331
Received 14 December 2024; Received in revised form 13 September 2025; Accepted 19 September 2025
Available online 10 October 2025
0010-0277/© 2025 The Author. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

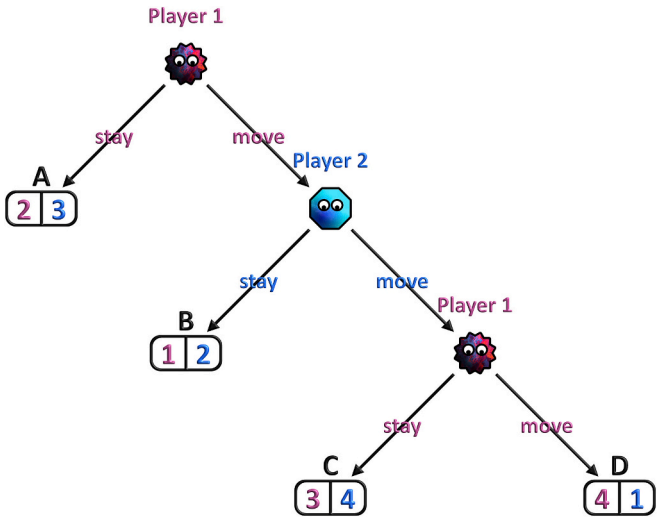


Fig. 1. Labeled example of a 3-step game: This figure illustrates a 3-step sequential-binary-choice game used in this experiment, showing the structure of player decisions and payoffs. Represented by avatars, choice nodes correspond to decisions made by players, while terminal nodes (leaf nodes) display payoffs for player 1 (left) and player 2 (right). Participants are always player 1. The game begins at the root node with player 1, depicted as the red and blue sun-shaped avatar. Player 1 can choose to stay or move. Staying ends the game at leaf node A, awarding player 1 a payoff of 2 and player 2 a payoff of 3. If player 1 moves, player 2, represented by the blue octagon, decides whether to stay or move. Staying ends the game at node B (1, 2), while moving gives player 1 the final decision between nodes C (3, 4) and D (4, 1).

root node, player 1 must reason backward through the game tree. Specifically, player 1 must predict what player 2 will think player 1 will do at the final node. For instance, in Fig. 1, at the final decision point, player 1 chooses node D because $C_1 < D_1$ ($3 < 4$). Anticipating this, player 2 compares B_2 and D_2 and prefers to stay at node B because $B_2 > D_2$ ($2 > 1$). Finally, player 1 compares A_1 and B_1, choosing to stay at node A because $A_1 > B_1$ ($2 > 1$). Such backward induction reasoning involves multiple nested levels of ToM: player 1 at the root node engages in level-2 (thinking about thinking about thinking), player 2 at level-1, and player 1 at the final node at level-0 (simple preference maximization). In these 3-step games, I classify level-2 choices at the root node as correct, and the opposite choices as errors.

While there is only one correct higher-order strategy for these games, multiple lower-order strategies are possible under cognitive strain. Out of a large model space, I focus on two archetypal extremes: the Blind Spot Model (BM) and Near Sight Model (NM). Both models assume hierarchical reasoning (0th, 1st, 2nd order, etc.) with deeper levels imposing more cognitive load. The BM proposes players under strain fail to represent some distant outcomes entirely ("blindness"), while the NM posits they approximate these distant outcomes probabilistically ("blurriness"; Fig. 2). This distinction yields testable predictions about how people cope when higher-order reasoning exceeds cognitive limits. Before detailing these models further, I first situate them within the broader family of *k*-level solution concepts, clarifying how these theories emerged as psychologically realistic alternatives to equilibrium-based game-theoretic reasoning.

1.1. Modeling ToM+ with psychologically plausible solution concepts

Classical game theory often relies on normative solution concepts: Nash Equilibrium (Nash, 1950; Nash, 1951) and rationalizability (Bernheim, 1984; Pearce, 1984). These approaches assume players reason about each other's strategies under the assumption of common knowledge of the game structure and of each other's rationality. This framework implies that players can take into account any arbitrary (potentially unbounded) order of beliefs, what one player thinks another player thinks, and so on, even though this is not a psychologically realistic description of human reasoning.

Crucially, higher-order Theory of Mind is both cognitively and computationally demanding: each added recursive step ("I think that you think that I think…") requires holding and manipulating ever-larger mental representations. If social reasoning is conceptualized as navigating a tree of possible states (as with game trees), the complexity can grow exponentially with each additional level of recursion when multiple future possibilities must be tracked.[2] As a result, humans are naturally limited in the depth of recursive reasoning they can achieve, motivating the need for psychologically plausible solution concepts—models that can explain not only how people succeed, but why they often fail at higher-order ToM tasks.

Behavioral Game Theory (BGT) addresses this gap by incorporating bounded rationality and human behavior observed in experimental games (Camerer, 2003). Within BGT, *k*-level models (Nagel, 1995; Stahl & Wilson, 1994; Stahl & Wilson, 1995) offer a more psychologically plausible account of strategic thinking by assuming players reason only to a finite depth. A level-1 reasoner anticipates the choices of a level-0 reasoner, a level-2 reasoner anticipates a level-1 reasoner's response, and so on.

In these models, level-0 players do not attribute beliefs or strategies to others. At terminal nodes, they simply maximize their own payoff. However, at choice points that lead to further moves by others, the level-0 player does not model those subsequent choices as intentional acts. The way that level-0 players treat the future choices of others varies by modeling assumption. For example,

- In some models, subsequent moves are treated as if made by chance, with no prior reason to favor one outcome over another—leading to uniform randomization consistent with the principle of insufficient reason (Keynes, 1921; Laplace, 1814).
- In other frameworks, level-0 players may assign greater likelihood to salient options or default choices, such as focal points (Schelling, 1960) or the status quo (Baron & Ritov, 1992), resulting in weighted averages rather than equal probabilities.
- At the most extreme, certain models assume that some possible game states are simply ignored or not represented at all, as in the "Blind Spot" Model.

The discrete hierarchy represented in basic *k*-level models is thus:

- 0th-order: I think
- 1st-order: I think you think
- 2nd-order: I think you think I think
- 3rd-order: I think you think I think you think, etc.

Cognitive Hierarchy Theory (CHT) (Camerer, Ho, & Chong, 2004) refines these ideas by allowing that players believe their opponents are drawn from a distribution over multiple levels below their own level. Unlike standard *k*-level models that treat the opponent as precisely one level lower, CHT posits that a level-*k* player assumes opponents are distributed across lower levels (e.g., level-0, level-1, level-2), typically following a Poisson distribution. This assumption introduces probabilistic reasoning about the population of opponents and provides a more flexible approach to modeling limited strategic thought.

Other descriptive models relax different assumptions. Social

[2] Formally, if each decision point branches into *b* future possibilities, then the set of possible states an agent must consider grows on the order of $O(b^k)$ for depth *k*. In the sequential binary choice games used in this study, the average branching factor is close to 1, so the growth is less dramatic, but the underlying principle remains: that recursive social reasoning is inherently tree-like and rapidly becomes computationally intensive.

projection theories (Krueger & Acevedo, 2005; Krueger, DiDonato, & Freestone, 2012) posit that players project their own reasoning onto others. The Noisy Introspection Model (Goeree & Holt, 2004) suggests that uncertainty about others' actions compounds with each level of reasoning. Alaoui and Penta (2016) propose an endogenous model where players weigh cognitive costs and benefits before deciding how deeply to reason. Each of these approaches acknowledges that people rarely achieve the high normative standards of game theoretic solution concepts but instead rely on heuristics and are bounded by information processing constraints.

1.2. Methodological challenges of inferring ToM level

Having outlined the broader theoretical landscape of *k*-level models, I now turn to the methodological challenge of diagnosing a participant's ToM level. To be clear, the current study measures participants' ToM levels, but it measures their ToM levels under time pressure specifically, rather than their ToM levels in general, as I will explain. In contrast to the sequential binary-choice games used here, many existing tasks either lack a clear sequence of recursive reasoning steps or rely on statistical inference to estimate ToM level, which is less precise than categorical classification.

1.3. Diagnosing ToM levels with games

While in theory, self-reporting could clarify participants' reasoning, such reports are unreliable in this context (Flavell, 1977). Therefore, researchers rely on structured tasks where behavioral choices signal reasoning depth. For instance, Perner (1979) pioneered the use of simultaneous binary-choice games to assess recursive reasoning in children. Simultaneous choices are defined as choices where all players commit to their choices without knowing others' decisions, even if moves are not literally simultaneous.

Payoffs within these games must be carefully structured to create a clear strategic starting point and avoid circularity. Consider Matching Pennies, a classic symmetric game with no evident starting point. Two players each place a penny face-up or face-down. If the coin faces match, player 1 wins both; if not, player 2 wins both. Because either action is equally risky and the payoffs are symmetric (± 1 for both players in all cases), no option stands out as an obvious choice and so there is no basis for beginning recursive reasoning.

Simultaneous discrete-choice games require asymmetric payoff structures to provide a clear cognitive starting point, as exemplified by the Nickel-Dime Game (Flavell, Botkin, Fry, & Jarvis, 1968; Perner & Wimmer, 1985). In this game, two players sit across from each other at a table with two cups turned over: one labeled "Nickel" covering a 5-cent coin, and the other labeled "Dime" covering a 10-cent coin. While player 2 turns around, player 1 silently removes one coin—either the nickel or the dime—and tosses it out of sight, leaving only one coin hidden under a cup. Then player 2 turns back and selects one cup to lift. If the chosen cup contains a coin, player 2 keeps it and player 1 gets nothing. If it's empty, player 1 receives the coin under the remaining cup. Because both players prefer the dime, the game provides a clear strategic base case: both initially aim for the higher value. However, this invites deception—player 1 may anticipate that player 2 will look under the dime-labeled cup and remove the dime to ensure a win. Yet player 2 might suspect this and go for the nickel instead, prompting player 1 to hide the nickel, and so on. As Perner (1979) explains, this leads to a recursive loop: even- and odd-level thinkers converge on identical choices, making it impossible to distinguish between players reasoning at levels 0, 2, 4,… *k*, or levels 1, 3, 5,… *k*+1, and thus rendering the game circular.

However, simultaneous discrete-choice games need not be circular if they are dominance-solvable. A dominated option is one that yields worse outcomes for a player regardless of the opponent's choice. A game is dominance-solvable if it can be fully reduced by iteratively eliminating dominated strategies, often alternating between players. The first eliminated option provides a natural starting point for recursive reasoning, and each subsequent elimination step can correspond to a level of Theory of Mind: a $k = 1$ player eliminates one dominated strategy, a $k = 2$ player eliminates two, and so on. Still, for this mapping to work the payoff structure must ensure that players at different reasoning depths make distinct choices, for instance, by placing a highly attractive payoff within an option that is eliminated only at higher levels.

Yet, dominance-solvable games do not always prescribe a unique cognitive path. Even if the game has a unique sequence of mathematically eliminable strategies—which is not always the case—participants may still focus first on eliminating options for player 1 or player 2. This flexibility introduces ambiguity about the order in which minds are represented, making it difficult to confidently map choices onto specific levels of Theory of Mind. By contrast, the *sequential* binary-choice games used in this study enforce a fixed mental order: players must reason about others in a strict series defined by the structure of the game tree.

Another approach comes from sequential continuous-choice games such as the *p*-beauty contest. In this task, *n* players each select a real number (e.g., between 0 and 1), and the winner is the one whose guess is closest to two-thirds the mean of all guesses. If others are choosing randomly, the expected mean is 0.5, so a $k = 1$ player would choose $3/9$. Anticipating that, a $k = 2$ player would choose $2/9$, and so on. As *k* increases, guesses progress downward toward 0—the Nash Equilibrium. This one-directional shift in behavior with increasing depth avoids the cycling problems found in many discrete-choice games and ensures that higher ToM levels yield distinct response patterns.

This makes the *p*-beauty contest useful for identifying recursive reasoning, but because responses are continuous, most guesses don't fall exactly on canonical level-*k* values. As a result, researchers must fit models to estimate each participant's likely ToM level. While this approach is valid and flexible, it introduces interpretive uncertainty compared to tasks like the one used here, where discrete choices allow for clean, categorical classification of reasoning depth on each trial.

1.4. Myopia-underestimation ambiguity

Regardless of the game type, it is worth noting a persistent challenge in higher-order Theory of Mind research is, what I call the myopia-underestimation ambiguity: when a player makes a choice consistent with lower-order reasoning, it is unclear whether this results from the player's own cognitive limitations (myopia) or from their belief that their opponent lacks higher-order reasoning capabilities (underestimation) (Alaoui & Penta, 2016; Goodie, Doshi, & Young, 2012). This ambiguity complicates the interpretation of behavior in *k*-level experiments, as observed behavior could reflect either self-imposed limitations or strategic adaptation to presumed opponent reasoning. However, in this paper I partially controlled for this ambiguity by consistently using a high-level opponent programmed to respond optimally at the maximum level of Theory of Mind, thereby giving participants an opportunity to learn the sophistication of their opponent and providing no indication of lower-order reasoning.

1.5. The Blind Spot and Near Sight Models

Building on these insights, the present study leverages a carefully

constructed set of sequential-choice games that can unambiguously diagnose ToM depth. Still, the ToM+ model determines the diagnostic criteria and so the diagnostic value of a game is only as valuable as the accuracy of the ToM+ model it uses. Finding an accurate ToM+ model can allow researchers to use these diagnostic games to study ToM+ and it's connections to other significant psychological phenomena, such as romance, gossip, sarcasm, deception, conflict, etc.

The current study tests two competing *k*-level models. These models, which I call the Blind Spot Model (BM) and the Near Sight Model (NM), differ only in their assumptions about how players reason at the very base case—0th-order reasoning—when their cognitive resources are exhausted and higher-level ToM+ is infeasible.

Both the BM and NM share the standard *k*-level structure at higher orders, assuming that fully rational, deeply recursive thinkers can hold the entire game tree in mind and use backward induction to predict outcomes. At these higher orders, both models make identical predictions: fully informed players choose optimally. However, recent evidence shows that when cognitive load forces people to abandon higher-order reasoning, they typically revert to the next lower ToM level (k - 1) rather than dropping randomly to any lower level (Arslan, Taatgen, & Verbrugge, 2017). The critical distinction between these models emerges only when reasoning depth is limited. Here, the BM and NM propose radically different heuristics for representing distant game states.

1.6. The Blind Spot Model (BM)

The BM posits that under cognitive depletion, players selectively disregard complex future states. In this baseline scenario, they effectively become "blind" to certain branches of the game tree. Inspired by Hedden and Zhang's (2002) 'ignore-D heuristic,' the BM aligns with the intuition that when complexity overwhelms cognitive capacity, humans prune their mental representations, focusing on only the nearest, most manageable nodes. Previous studies have also employed or discussed the 'ignore-D heuristic' (Flobbe, Verbrugge, Hendriks, & Krämer, 2008; Goodie et al., 2012; Meijering, Rijn, Taatgen, & Verbrugge, 2011; Meijering, Taatgen, Rijn, & Verbrugge, 2014; Meijering, Van Rijn, Taatgen, & Verbrugge, 2012; Zhang, Hedden, & Chia, 2012; Zhen & Yu, 2021).[3] Originally, Hedden and Zhang's spatial representation of games (with cells A-B-C-D arranged on a board) may have encouraged focusing on the "farthest" cell, D, as the hardest to represent. Yet, in non-spatial game trees, distance is found by counting nodes, not measuring inches, making the BM's choice of which node to ignore appear somewhat arbitrary. Still, the BM's essence remains a compelling idea: under strain, some players might just omit certain possibilities entirely, as if using selective attention (Broadbent, 1958; Pashler, 1998). Likewise, cognitive models suggest that humans prune selectively during multi-step planning—e.g., skipping branches containing large losses or unlikely payoffs—consistent with resource-rational trade-offs between expected gain and computation (Huys et al., 2012; Lieder & Griffith, 2020).

Additionally, the idea of selectively ignoring information has parallels in computer science, where algorithms routinely "prune" branches when exhaustive search becomes computationally infeasible. For example, chess programs use alpha-beta pruning to eliminate moves that cannot improve the best option found so far; search algorithms use branch-and-bound to cut paths that cannot beat a known solution; and Monte Carlo Tree Search methods focus computational resources on the most promising options while starving less valuable branches of attention (Browne et al., 2012; Knuth & Moore, 1975; Kocsis & Szepesvári, 2006; Land & Doig, 1960; Newell, Shaw, & Simon, 1958; Pearl, 1984; Russell & Norvig, 2010). These algorithmic approaches typically use sophisticated criteria to decide what to ignore, intelligently pruning based on which branches are least likely to matter. The BM represents a simplified version of this general strategy: rather than using complex criteria, it simply eliminates certain future possibilities when cognitive resources are strained.

1.7. The Near Sight Model (NM)

By contrast, the NM assumes that even when reasoning depth is limited, players do not completely ignore distant states. Instead, they approximate these future outcomes probabilistically, treating a complex choice node as if it were a "chance node" with equal probabilities assigned to each unreachable but relevant outcome. Treating choice nodes like chance nodes incorporates the fact that social reasoning is more computationally complex than nonsocial reasoning when people's choices are strategically interdependent (von Neumann & Morgenstern, 1947). Moreover, this fuzzy averaging embodies bounded rationality (Simon, 1955; Simon, 1957). Rather than discarding distant options, the NM suggests humans maintain a simplified representation of all outcomes, just in a blurred, probabilistic form. This ensures no node is arbitrarily discarded and reflects a psychologically plausible coping strategy: instead of all-or-nothing, the NM settles for "good enough" approximations under cognitive load.

Rather than pruning subtrees, a complementary and widely used strategy under tight computational budgets is to summarize difficult-to-predict parts of a decision tree with probabilistic defaults rather than to model them exhaustively. In metareasoning and resource-rational analysis, agents are predicted to stop costly look-ahead when the expected value of additional computation is low and to act using quick summaries of the remaining uncertainty (e.g., expected values under simple priors). This yields noisy, "good-enough" choices that trade accuracy against cost (Gershman, Horvitz, & Tenenbaum, 2015; Lieder & Griffith, 2020; Russell & Wefald, 1991). In planning under uncertainty, exact belief-state reasoning over other agents becomes rapidly intractable (POMDPs/I-POMDPs), so approximate planners replace deep opponent modeling with sampling or default distributions over future outcomes and back up expected utilities instead of fully resolved choices (Doshi & Gmytrasiewicz, 2009; Kaelbling, Littman, & Cassandra, 1998). A parallel psychological precedent comes from Bayesian accounts of perception, where when evidence quality drops, priors dominate and the brain fills in missing detail with generic expectations (Knill & Pouget, 2004; Weiss, Simoncelli, & Adelson, 2002). The NM formalizes this gap-filling strategy for ToM+: when predicting an agent's choice becomes too demanding, it treats unresolved choice points as chance nodes with default probabilities (e.g., equal weights).[4]

1.8. A spectrum of base-case heuristics

The BM and NM thus represent two endpoints on a spectrum of possible base-case heuristics. The BM uses an extreme "hard cutoff" approach (0 or 1 weighting of nodes), while the NM adopts a uniform averaging (equal weighting) approach. Yet, one can imagine intermediate base cases. For instance, focal-point reasoning could allow players to assign nonuniform probabilities based on salience (Schelling, 1960) or default outcomes (Baron & Ritov, 1992). In scenarios where one future node is especially prominent—e.g., it offers a standout payoff, is visually salient, or is recognized as a default outcome if the opponent fails to act—a player might weight the probability of that node more heavily than others. This focal-point base case would rest between the

[3] Using Hedden and Zhang's (2002) payoff structures to discriminate between ToM level is equivalent to relying on the 'ignore-D heuristic'.

[4] The NM uses equal probabilities and multiplies them by the payoffs to compute expected payoffs, which involves traversing the child nodes. An even simpler strategy would be to guess expected payoffs from a learned baseline, such as 2.5, which is the mean of the payoffs used this experiment {1, 2, 3, 4}. I simulate a model that uses this strategy in the results.

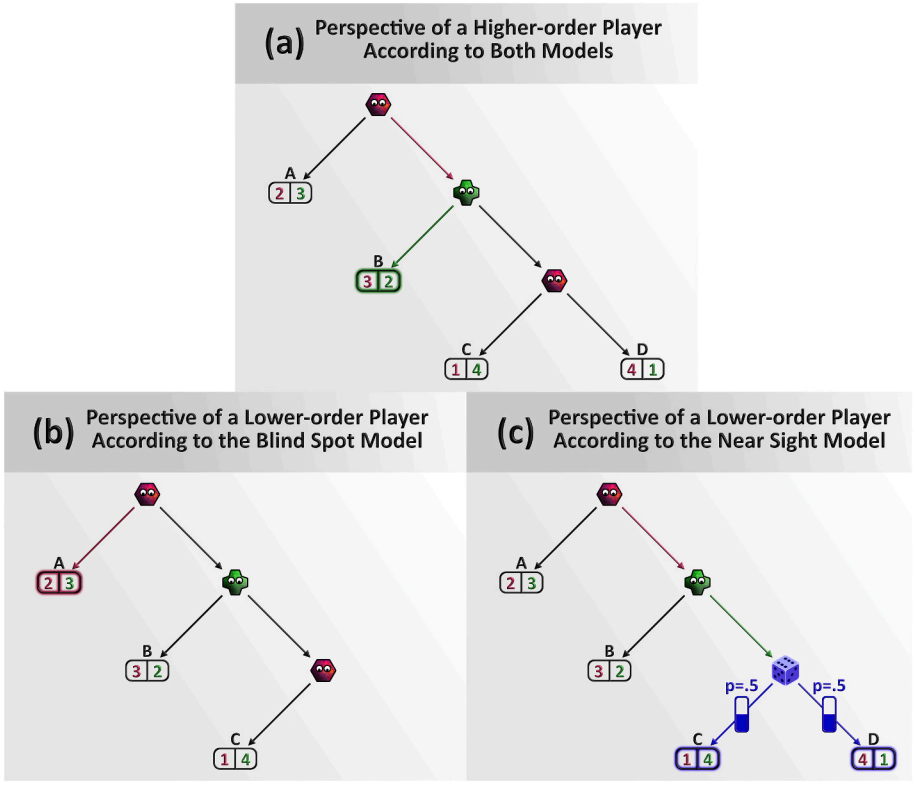


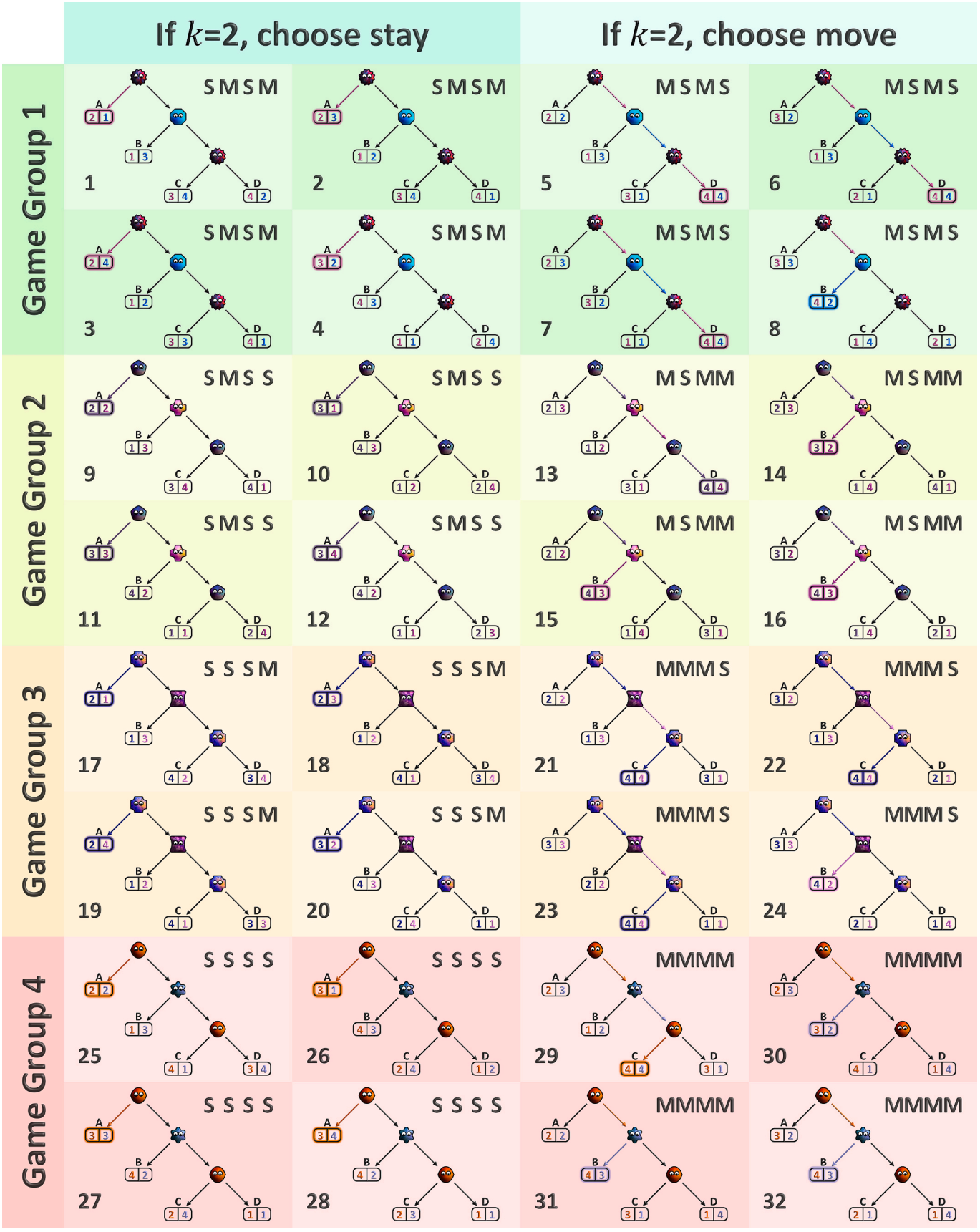
Fig. 2. Mental representations of games: As shown in panel (a), the BM and NM agree that a higher-order player's mental representation of a game is complete but use different heuristics to explain failures of ToM+, each positing mental representations that are incomplete in different ways. In panel (b), the BM omits an entire node, focusing only on the first child node, somehow assumed to be *nearer* to the player. In panel (c), the NM omits a choice node, replacing it with a chance node, assuming that it is more cognitively demanding to predict intentional agents than non-mental systems.
Each model predicts a different outcome when considering the game from the perspective of player 1 at the root node: in cell (a), from a higher-order perspective, both models expect the game to end at node B; in cell (b), from a lower-order perspective, the BM expects the game to end at node A; and in cell (c), from a lower-order perspective, the NM expects the game to end at node C or D, depending on a chance event. These final outcomes, and paths leading to them, are colored for visual clarity (as they would be on the user interface at the end of a round).

BM's all-or-nothing blindness and the NM's uniform averaging. Thus, the BM and NM are anchors at opposite ends of a broader continuum, outlining the range of possible strategies that depleted cognition might use to represent future states.

1.9. Example predictions of both models

Fig. 2 is an example of how both models expect the same choices from higher-level players but different choices from lower-level players. The payoffs are: A (2, 3), B (3, 2), C (1, 4), and D (4, 1). Both models expect a higher-level player 1 to work backward from the final decision point—thinking about what player 2 will think they (player 1) will choose. Comparing C (1, 4) to D (4, 1), player 1 will choose D because $C_1 < D_1$ ($1 < 4$). Given that player 1 will choose D and D_2 is only a payoff of 1, player 2 will choose to stay because $B_2 > D_2$ ($2 > 1$). Realizing that player 2 will choose end the game at node B and that $A_1 < B_1$ ($2 < 3$), player 1 will choose to move at the root node. However, according to both models, when deciding to move or stay at the root node, a lower-order player 1 can represent how player 2 will represent their choice but will fail to fully represent how player 2 will represent their (player 1's) final choice.

According to the Blind Spot Model (BM), node D is entirely omitted from a lower-order player 1's mental representation. From this lower-order perspective, player 1 has no final choice and so player 2 only compares $B_2 < C_2$ ($2 < 4$) and so will choose to move if given the chance, resulting in player 1 receiving payoff C_1, which is only 1. Preferring A_1 to C_1 because ($2 > 1$), player 1 chooses to stay at the root node. Because the BM predicts that a lower- and higher-order player will make opposite root node choices, the BM classifies this as a *diagnostic* game because the root node choice is diagnostic of ToM level. By contrast, according to the Near Sight Model (NM), the payoffs nodes C and D are averaged, as if a lower-order player 1 imagines that player 2 imagines that the final node is a chance node with a 0.5 probability of leading to node C and a 0.5 probability of leading to node D. This yields an expected payoff of (2.5, 2.5). Comparing B_2 to this probabilistic value, player 2 prefers to move because $2 < 2.5$. Anticipating player 2's move decision, player 1 compares A_1 to their expected payoff and chooses to move because $2 < 2.5$. Thus, the NM predicts that a lower- and higher-order player will make the same root node choice, the NM classifies this as a *nondiagnostic* game because the root node choice cannot distinguish lower- and higher-order players. As I will explain, it is the fact that the models disagree about which games are diagnostic, combined with the fact that participants are



(caption on next page)

Fig. 3. All 32 payoff structures – categorization and counterbalancing: Panels show all 32 payoff structures within game trees, categorized by game groups: 1) both models classify these games as diagnostic and thus expect high error rates, green, 2) only the BM classifies these games as diagnostic and thus expects a high error rate, yellow, 3) only the NM classifies these games as diagnostic and thus expects a high error rate, orange, and 4) neither model classifies these games as diagnostic and so neither model expects high error rates, red. The bottom-left number is the game index (1−32). The top-right code is the quadruplet (e.g., *SMSM, MSMS,* …), which encodes each model's root action at two depths in the order [BM $k = 2$, BM $k = 1$, NM $k = 2$, NM $k = 1$] with S = stay and M = move. Each game group contains 4 games where the optimal higher-order root node choice is to 'stay', shown under "If $k = 2$, choose stay" on the left and 4 games where the optimal higher-order root node choice is to 'move', shown under "If $k = 2$, choose move". Payoffs and choices match the color of their respective avatar. The path from the root node to the terminal node is colored to indicate where the game ends if both players choose optimally. By design, all stay-leading games terminate at A; among move-leading games, 8 terminate at B, 4 at C, and 4 at D. The set of 32 payoff tuples was counterbalanced for (*i*) starting payoff $A_1 \in \{2, 3\}$, (*ii*) the up–down payoff trend across levels, and (*iii*) whether the optimal root action was move or stay, to prevent reliance on simple associations (e.g., "move if $A_1 = 2$"). Exact tuples and the up–down factors are listed in Table 4.

more likely to make mistakes in diagnostic games, that enables this study to test which of these models best explains ToM+ heuristics.

2. Methods

This experiment tested the Blind Spot Model (BM) against the Near Sight Model (NM) to see which model, if any, could predict participant error rates across a set of 32 alternating-choice games with systematically varied arrangements of payoffs, called payoff structures:

1. Eight games where both models expect a higher proportion of mistakes,
2. Eight games where the BM expects a higher proportion of mistakes,
3. Eight games where the NM expects a higher proportion of mistakes,
4. Eight games where neither model expects a higher proportion of mistakes.

Shown in Fig. 3 (and Table 4 in the Appendix), these games are counterbalanced by the following features: (1) if the optimal root-node choice is to stay or to move, (2) if payoff A_1 is 2 or 3, called 2-starting or 3-starting games by Hedden and Zhang, (3) the pattern of player 1 payoffs increasing or decreasing from nodes A to B to (C and D), conceived of by Hedden and Zhang (2002), and even (4) the sum of all the A_1, B_1, C_1, D_1, A_2, B_2, C_2, and D_2 payoffs across all 32 games is exactly 80, creating mathematical symmetry. The task does not bias participants to learn to associate specific payoff positions with higher rewards because the mean payoff in each position is 2.5 (80/32).[5]

These are 2-player, binary, sequential choice games where players cannot communicate, and where the rules are known to be mutually understood. Choices always follow the same sequence: player 1 (participant), player 2 (artificial agent), and player 1 (participant). There are four possible ending states: A, B, C, and D, each with rewards called payoffs for both players. Payoffs are always drawn from the set {1, 2, 3, 4} without replacement. To eliminate the confounding influences of social preferences like altruism and fairness and strategic reasoning across games, participants are instructed to behave selfishly, to maximize their payoff one game at a time.

Figure 4 below provides example games to illustrate how choices were classified and how payoffs structures were manipulated to produce the four game categories. Each model expects a specific choice to be made by a 2nd-order (higher-order) player and a specific choice to be made by a 1st-order (lower-order) player. A game is classified as diagnostic if the 2nd- and 1st-order choices at the root node are opposite, and as nondiagnostic when players at both levels will make the same choice. This classification does not depend on whether the expected 2nd- or 1st-order choice was "stay" or "move" and these choice types were equally distributed within each of the four game types.

Diagnosticity for the NM arises at the final choice when player 2's payoff (B_2) lies between the average of outcomes at subsequent endpoints (C_2 and D_2) and their likely payoff if player 1 selects either C or D. Likewise, diagnosticity for the BM arises when player 2's payoff (B_2) lies between C_2 (with D_2 omitted from represented) and their likely payoff if player 1 selects either C or D. Accurate prediction at the root thus requires player 1 to reason about player 2's beliefs regarding player 1's own future choice. If player 1 cannot represent this level of recursion, the resulting choice at the root can differ from the optimal higher-order strategy.

Both the BM and the NM expect 2nd-order players to make the same root node choices, but they can disagree about the choices of 1st-order players. Because they disagree about 1st-order choices, they can disagree about which games are diagnostic or nondiagnostic, depending on the payoff structures. A mistake is failing to make the optimal 2nd-order choice at the root node and participants are expected to make more mistakes on diagnostic games than on nondiagnostic games because in diagnostic games, players must overcome lower-order thinking to decide correctly. By contrast, in order to choose incorrectly at the root node in nondiagnostic games, one would have to contradict the best response of a 2nd- *and* 1st-order player, making such an event much less likely. Because the two models can disagree about which games are diagnostic and players are expected to have higher error rates on diagnostic games, I varied the payoff structures to create the four groups of games: (1) games 1–8, which both models classify as diagnostic and so expect high error rates, (2) games 9–16, which only the BM classifies as diagnostic and thus expects a high error rate, (3) games 17–24, which only the NM classifies as diagnostic and thus expects a high error rate, and (4) games 25–32, which neither model classifies and diagnostic nor expect a high error rate.

While I focus on a version of the BM where 1st-order reasoners ablate node D, based on the 'ignore-D heuristic' (Hedden & Zhang, 2002), I could have also focused on a version of the BM where 1st-order reasoners ablate node C, based on an opposite 'ignore-C heuristic'. The payoff structures are counterbalanced between games such that the ignore-D BM expects errors only in game groups 1 and 2, while the ignore-C BM expects errors only in game groups 3 and 4. A random BM that ignores node C or node D with equal probability would expect errors to be uniformly distributed across game groups. This experiment is equally a test of these versions of the BM.

Also, one might wonder whether an observed error pattern could be explained not by different heuristics, but simply by noise in higher-order reasoning. For instance, if participants attempted second-order ($k = 2$) reasoning in every game but occasionally made random mistakes by flipping their intended choice, could this produce systematic differences in error rates between groups?

However, the counterbalancing of the payoff structures ensures that a model of higher-order reasoning with decision noise would produce error rates that would converge toward a uniform distribution across game groups. Both models predict exactly the same second-order

[5] Model-based vs. model-free reinforcement-learning tasks, such as the Daw, Gershman, Seymour, Dayan, and Dolan "two-step" paradigm (2011), dissociate choices that rely on a learned state transition model from choices that track average rewards historically associated with an action across trials. In the present task, the full transition structure and all payoffs are visible on every trial, and the payoff set was counterbalanced so that the mean payoff in each position is exactly 2.5. Consequently, position-based value learning would predict no preference for *stay* or *move*. Any systematic differences in errors therefore reflect evaluations of consequences (who benefits at C vs. D), not model-free value learning.

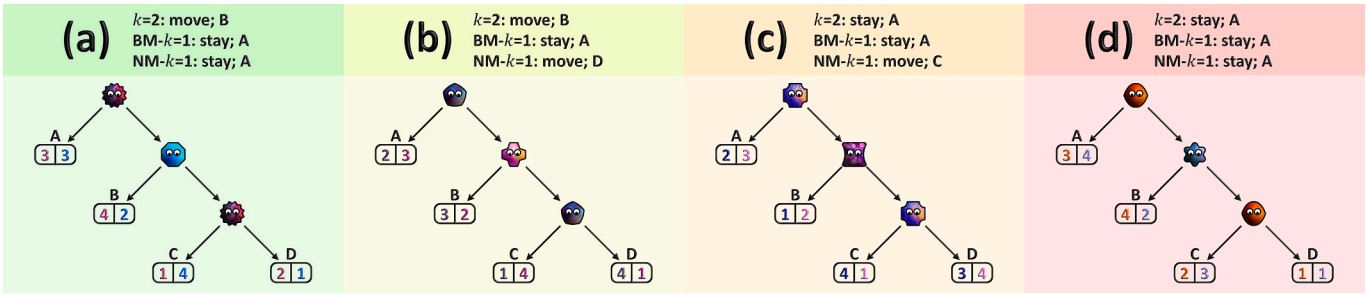


Fig. 4. Example games illustrating diagnostic categories: Games are divided into four groups according to whether the Blind Spot Model (BM) and the Near Sight Model (NM) classify them as diagnostic or nondiagnostic of Theory of Mind (ToM) level: (a) game group 1: diagnostic for both models, (b) game group 2: diagnostic for BM only, (c) game group 3: diagnostic for NM only, and (d) game group 4: nondiagnostic for both. A game is considered diagnostic when the root node choices predicted for higher-order ($k = 2$) and lower-order ($k = 1$) players are opposite (stay vs. move). Models expect higher error rates on these diagnostic games, as optimal responding requires overcoming the simpler logic of lower-order reasoning. Errors are defined as deviations from the higher-order ($k = 2$) choice.
The top of each panel displays the predicted root node choice for a $k = 2$ player (agreed upon by both models), for a $k = 1$ player according to BM, and for a $k = 1$ player according to NM. Letters indicate the expected final node.

choices across all games. Furthermore, as shown in Fig. 3 (and Table 4 in the Appendix), each game group contains an equal number of games where the correct second-order response is "stay" and "move," ensuring no directional bias. If a participant were simply adding noise to a consistent second-order policy, their mistakes would be evenly distributed across game groups.

The choice order always followed: player 1, player 2, player 1, if the game did not end after the first or second choice. I required participants to respond within 8 s to ensure that mistakes occurred. Mistakes are deviations from optimal second-order play at the root node. The *Morality Game* has a countdown timer that shows participants the remaining time to respond before their response is 'abdicated', resulting in a payoff deduction, which was a deduction of 10 payoffs in this experiment. This time pressure was intended to be short enough to ensure that many participants would not be able to engage in complete backward induction much of the time but long enough to ensure that responses were not random.

The choice of an 8-s time limit was informed by pilot testing and prediction reaction times in Hedden and Zhang (2002). Participants in pilot tests reported that an 8-s window struck a balance between allowing for deliberate, non-random responses and inducing sufficient time pressure to prevent exhaustive recursive reasoning in most cases. Crucially, the resulting error patterns across game groups in this experiment demonstrate that the time constraint was well-calibrated: error rates were systematically higher in the groups predicted by theory, indicating engagement with the task rather than random guessing or floor effects. If the time limit had been overly generous, errors would have been rare; if too brief, errors would have been indiscriminately distributed. Time pressure is a well-recognized method for inducing cognitive load and encouraging reliance on heuristics (Payne, Bettman, & Johnson, 1993). Under limited time, individuals typically resort to simplified processing strategies, such as focusing on the most salient

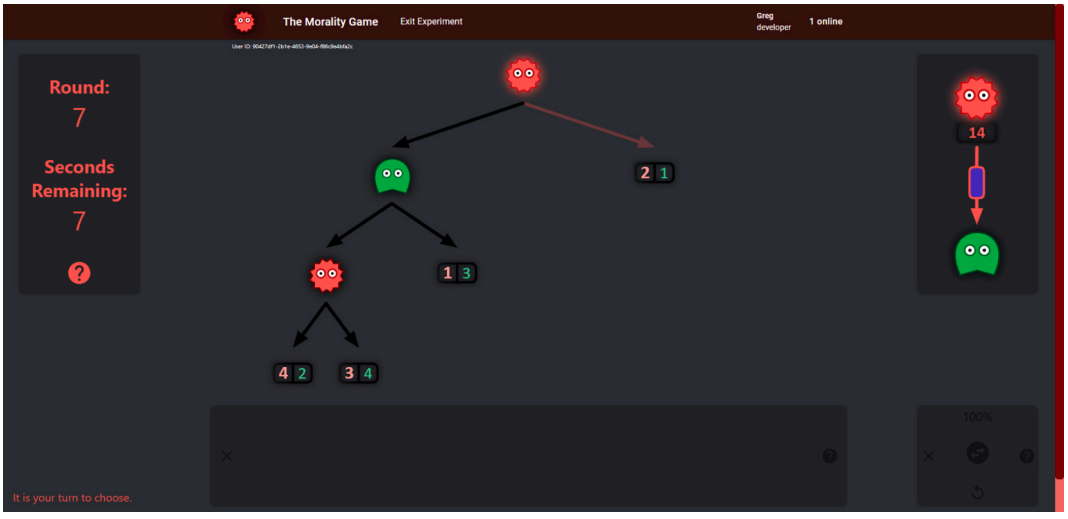


Fig. 5. Example morality game user interface: This shows the user interface on my new *Morality Game* platform (https://moralitygame.com). The participant's avatar is light red, and their counterpart's avatar is green. On round 7, with 7 s remaining, the participant is about to select a payoff of 2 for themselves and a payoff of 1 for their counterpart. Participants scroll through options via keyboard buttons "A" (left) or "D" (right) or by dragging their mouse or finger (on touchpads) over the darkened rectangle on the bottom of the screen. The interior of the top left arrow is colored light red to indicate which option the participant will select if they submit their choice by pressing the spacebar, clicking the mouse, or tapping the touchpad. Once submitted, the UI confirms their choice by darkening the arrow and emitting a sound.

information or employing probabilistic reasoning, rather than conducting comprehensive analyses. This method assumes that time pressure would cause similar heuristics as other forms of cognitive load or depletion, such as sleep deprivation and multitasking.

The winner of the contest is the model that correctly identifies the high-error game groups in my planned X^2 contrasts (Bonferroni-adjusted $\alpha = 0.0125$).

- If more mistakes are made in game groups 1 and 3, shown above, then the NM wins.
- If more mistakes are made in game groups 1 and 2, then the BM wins.
- Otherwise, neither model wins.

2.1. Participants and apparatus

84 participants were recruited from the University of Michigan Psychology Subject Pool using SONA Systems and participated for class credit. The experiment was conducted remotely via Zoom, with all participants engaging simultaneously. The *Morality Game* automatically ejects inactive participants to prevent them from stalling experiments. 10 participants were excluded due to inactivity (e.g. being idle, connectivity issues, switching browser tabs). One participant was excluded for abdicating on all 32 root choices (never making a choice in time). Thus analyses used $N = 73$ participants. Abdications are not scored as errors and so cannot influence the distribution of error rates across game groups. Table 5 in the Appendix shows that 66/73 participants abdicated 3 times or fewer. This study utilized the *Morality Game* platform https://moralitygame.com, where each participant was required to undergo a standard tutorial and successfully answer six comprehension questions to ensure understanding of the game mechanics (see Fig. 5).

To familiarize participants with the gameplay, each engaged in a minimum of six practice rounds against live human opponents, featuring a variety of game trees. This session was crucial for acclimatizing participants to the multiplayer experience. Participants joined a Zoom call in groups of 6–12 so that I could help them navigate the website, enter the tutorial, and join the experiment. Playing these practice rounds against human participants demonstrates that the platform can host live peer-to-peer interactions and reduces the chances that they would suspect that their counterpart was actually a bot. They were told that they would be matched with the same human counterpart across all 32 rounds.

All participants encountered the same 32 sequential-choice games but the order of games was randomized for each participant. In these trials, all participants assumed the role of player 1, with a consistent 8-s response window at the root node and competed against a bot programmed to respond optimally at the maximum level of Theory of Mind. These optimal preprogrammed responses avoided biasing participants in favor of the heuristics of either theoretical model under investigation, instead encouraging higher-level thinking. The maximum ToM level for the bot-counterpart was only level 1 because the bot was always player 2, and there is only one more choice point that can follow player 2's decision, meaning only one more mind to consider.

Table 1
Game payoff structures are categorized by whether the BM and NM expect high or low error rates.

Group Number	Game Indices	BM expected error rate	NM expected error rate
1	1–8	High	High
2	9–16	High	Low
3	17–24	Low	High
4	25–32	Low	Low

As is standard for any experiment run on the *Morality Game* platform, when the game reaches a decision node, players not assigned to the role of chooser must submit a prediction for what choice they expect the chooser to make. These predictions are binary and are submitted just like choices. While prediction data was not used as evidence for or against either model, the fact that participants were made to submit predictions prevented them from seeing the immediate reaction times of their computerized counterparts because the system hides the choices from predictors until after their predictions are made. This means that participants could only surmise that their counterparts submitted responses at least a bit faster than they did and only in the games where they chose "move" at the root node, allowing player 2 to choose at the next choice point, which occurred in roughly half of all games.[6]

The visual presentation of the game was carefully controlled; participant and counterpart avatars varied randomly in color and shape, while the sibling node order within the game trees was standardized to facilitate specific conditions favorable to the Blind Spot Model by always placing node C left of node D.

Data capture was automated by the *Morality Game* platform, which recorded all interactions within the games and stored the results in two formats: a JSON file listing game trees with corresponding participant responses, and a CSV file structured for efficient subsequent analysis.

2.2. Analysis

The analysis was conducted by comparing the number of correct and mistaken root node choices across the four game categories. Error rates were calculated as the proportion of incorrect responses out of total responses (incorrect + correct). To ensure robust estimates of the error rates and their uncertainty, 95 % confidence intervals were computed using a bootstrapping approach with 10,000 resamples. This technique involved generating repeated random samples, with replacement, from the observed response data, and calculating the error rate for each resample. The 95 % confidence interval was derived by taking the 2.5th and 97.5th percentiles of the bootstrapped error rate distribution.

Participants failed to respond within the 8-s window in 4.63 % of all root node choices (108/2336). I did not count these abdications as errors. Table 2 shows the number of abdications in each game group. Table 6 shows that treating abdications as errors would not have substantially altered results.

Chi-square tests were employed to examine whether there were statistically significant differences between the groups' error counts, as predicted by the Blind Spot Model (BM) and the Near Sight Model (NM). Specifically, I conducted the following four comparisons (see Table 1):

- **(1&3) vs (2&4):** To test the NM's prediction that error rates are higher in groups 1 and 3.
- **(1&2) vs (3&4):** To test the BM's prediction that error rates are higher in groups 1 and 2.
- **1 vs 3:** To compare error rates within the high-error groups predicted by the NM.
- **2 vs 4:** To compare error rates within the low-error groups predicted by the NM.

I applied a Bonferroni correction to control for the increased risk of Type I errors due to multiple comparisons. With four comparisons, the adjusted alpha level was set to: $\alpha = \frac{0.05}{4} = 0.0125$.

3. Results

The results decisively supported the Near Sight Model (NM), which

[6] It is plausible that player 2 (counterpart) would respond faster than player 1 (participant) because player 2 can spend extra time analyzing their choice while player 1 is making their decision at the root node.

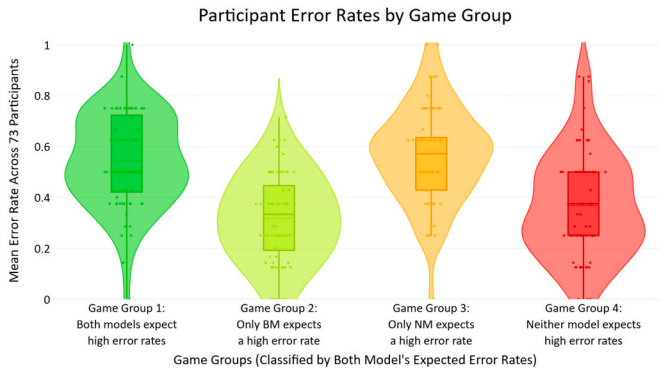


Fig. 6. Error rates by game group: This violin chart presents this paper's central result, demonstrating the success of the Near Sight Model in predicting which games participants would make the most errors in, in contrast to the Blind Spot Model's inability to predict errors. Both models expect more errors to occur in games where higher- and lower-order players make opposite choices at the root node, called diagnostic games, because in such games optimal play requires overcoming lower-order ToM reasoning. However, these models disagree about which games are diagnostic versus nondiagnostic and thus more versus less likely to cause mistakes. The decisive evidence favoring the NM over the BM is that participant error rates corresponded to the NM's criteria for classifying games, not the BM's criteria, as shown by the higher mean error rates in the first and third game groups versus the lower mean error rates in the second and fourth game groups. The dots within the violins are individual participant error rates, which tend to fall in regular intervals because error rates are *n* errors divided by the total number of choices per game group, where the total is always 8 minus abdications. The green-yellow-orange-red color scheme applies to game groups throughout the paper.

accurately predicted the error rates across the four game categories (see Fig. 6). In contrast, the Blind Spot Model (BM) did not significantly predict the outcomes, thus challenging its heuristic assumptions about player behavior. This finding aligns with the NM, which suggests that lower-order players average over distant game states and effectively falsifies the notion that participants consistently overlook entire game states. Likewise, these results falsified the mirror version of the BM that ablates node C, which expects errors only in game groups 3 and 4, and the random BM that ablates node C or node D with equal probability and expects errors to be uniformly distributed across game groups. The sharp differences in error rates also falsify the model of higher-order reasoning with decision noise, which would predict uniformly distributed errors.

3.1. Statistical analysis

- **(1&3) vs (2&4):** The chi-square test comparing groups 1 and 3 (high-error categories predicted by the Near Sight Model, NM) to groups 2 and 4 revealed a highly significant difference, $\chi^2(1, N = 2228) = 104.99, p < 0.0001$. The effect size, measured by Cramér's $V = 0.217$, indicates a small-to-medium association. This finding aligns strongly with the NM's predictions and remains robust after applying the Bonferroni correction ($\alpha = 0.0125$). The large discrepancy in error rates (0.55 in groups 1 and 3 versus 0.35 in groups 2 and 4) shows the practical significance of these results.
- **(1&2) vs (3&4):** The chi-square test comparing groups 1 and 2 to groups 3 and 4 did not yield a significant difference, $\chi^2(1, N = 2228) = 2.523, p = 0.112, V = 0.034$. The lack of statistical significance and the very small effect size provide no support for either version of the Blind Spot Model's predictions, where the ignore-C BM predicts higher error rates in groups 3 and 4 and the ignore-D BM predicts higher error rates in groups 1 and 2.
- **1 vs 3:** Within the high-error categories predicted by the NM (groups 1 and 3), no significant difference was observed, $\chi^2(1, N = 1110) = 0.466, p = 0.495, V = 0.021$. This result confirms that the NM accurately predicts consistent error patterns across these groups.
- **2 vs 4:** Similarly, within the low-error categories predicted by the NM (groups 2 and 4), no significant difference was found, $\chi^2(1, N = 1118) = 3.177, p = 0.075, V = 0.053$. While this result approaches significance, the small effect size suggests minimal practical differences between these groups.
- **Abdications:** Abdication rates across the four groups were not significantly different, $\chi^2(1, N = 2336) = 11.26, p = 0.010, V = 0.069$. The uniformity of abdication rates further supports the robustness of the NM predictions, as this potential confound was distributed evenly. Since this comparison is exploratory and does not directly test my main hypotheses, I did not include it in the Bonferroni correction.
- **Omnibus Uniformity:** An omnibus chi-square test across all four game groups decisively rejected the null hypothesis that error rates were uniformly distributed, $\chi^2(3, N = 2228) = 109.51, p < 0.0001$, $V = 0.222$. This rules out models that predict evenly spread errors, including the random Blind Spot Model and the model of higher-order reasoning with root node decision noise.

Additionally, I compared the ratio of errors between game groups 2 and 3 at the individual level to assess whether a subpopulation of participants were best described by the BM despite that the NM better predicted the distribution of errors at the group level. Because only the BM expects high error rates in game group 2 and only the NM expects high error rates in game group 3, the ratio of errors made in game group 2 to the total number of errors made in game groups 2 and 3 tracks the degree to which the BM best predicts that participant's errors. I excluded 4 participants from this analysis with error counts too low to form meaningful ratios (2 or fewer mistakes in game groups 2 and 3). 9 participants had ratios of 0.5, meaning that both models predict their errors equally well. The NM best predicted the error rates of 53 participants and the BM best predicted the error rates of 5 participants. Of these 5 participants, all had error ratios under 0.7. This does not strongly suggest that the BM best characterizes a subpopulation of participants.

Table 2
Error count across four game categories [310, 175, 310, 207]: The Blind Spot Model expects a disproportionately high error rate in the *first* and *second* game categories because the BM classifies those games a diagnostic. The Near Sight Model expects a disproportionately high error rate in the *first* and *third* game categories because the NM classifies those games a diagnostic. The error count/rate is high only in the first and third game categories, as predicted by the NM. The 95 % confidence intervals show overlaps within the first and third, and second and fourth categories respectively, but not between the high-error and low-error groups as predicted by the NM. This pattern substantiates the NM's diagnostic accuracy in predicting player errors based on game category.

BM Expects More Errors	NM Expects More Errors	Correct Count	Error Count	Abdicate Count	Error Rate	95 % CI Lower	95 % CI Upper
TRUE	TRUE	256	310	18	0.54805	0.50707	0.58834
TRUE	FALSE	380	175	29	0.31653	0.27748	0.35495
FALSE	TRUE	234	310	40	0.56985	0.52757	0.61213
FALSE	FALSE	356	207	21	0.36935	0.32860	0.40675

Table 3
How randomized node C/D payoffs distribute errors across game groups: I simulated a Random Defaults Model (RDM) that replaces payoffs from nodes C and D with randomly sampled values from {1, 2, 3, 4} to see if it could reproduce the observed distribution of errors of [0.3094, 0.1747, 0.3094, 0.2066] in game groups [G1, G2, G3, G4], shown on the bottom row. Each G1-G4 value is the number of errors produced by the RDM in a game group divided by the total number of errors across all game groups. I also simulated a Random Transpose Model (RTM) that transposes pairs of payoffs from nodes C and D. These payoff alterations symbolize forms of corruption to a lower-order player's mental representation of the game. Iterating 1,000,000 times over all 32 games, I simulated multiple versions of each model by varying which payoffs were randomized or transposed, shown on each row. Both models show that the observed error pattern of (G1 & G3) > (G2 & G4) can be best approximated by distributing noise (payoff alteration) across nodes, than localizing it within single nodes, akin to what the NM does.

Payoffs Randomized/Transposed				Simulated Game Group Error Proportions				Model Fits	
C_1	C_2	D_1	D_2	G1	G2	G3	G4	RDM	RTM
True	True	False	False	0.0422	0.0840	0.5000	0.3738	0.2924	0.5314
False	False	True	True	0.4580	0.2896	0.1262	0.1262	0.4648	0.0400
True	False	True	False	0.2384	0.1914	0.3008	0.2694	0.8502	0.9134
False	True	False	True	0.3333	0.1666	0.3333	0.1666	0.8947	0.9134
True	True	True	True	0.2949	0.2051	0.2949	0.2051	0.9317	–
Participant Error Proportions:				0.3094	0.1747	0.3094	0.2066	1.0000	1.0000

Fig. 8 shows the error ratio distribution in the Appendix.

Median root-node reaction times did not significantly differ between correct and mistaken choices (Wilcoxon $W = 1186$, $p = 0.336$, $N = 73$; medians 3.92 s vs 3.78 s) (Fig. 9). Median root-node reaction times varied only modestly across the four game groups (Friedman $\chi^2(3) = 6.584$, $p = 0.086$). No pairwise contrast survived Bonferroni correction, but the largest uncorrected difference was G2 > G3 (raw $p = 0.035$, corrected $p = 0.213$). Thus, the payoff conditions had a far stronger effect on error rates than on reaction times.

Additionally, the primary analysis focuses on root node choices, not predictions, because participants only had the opportunity to submit predictions in roughly half of the games where they chose to "move" because choosing to "stay" ends each game. Although choosing at the root node requires representing a larger game tree than predicting at the second step of the game, predictions still require 2nd-order reasoning: thinking about player 2 thinking about player 1's final choice. A supplementary analysis shows that participants tended to make predictions at the same ToM level as their previous choice 77.6 % of the time (Table 7; Appendix). This response consistency seems to suggest that participants employed coherent and persistent ToM reasoning with each game.

3.2. Simulating alternative models

The Blind Spot Model (BM) and the Near Sight Model (NM) are archetypes—idealized endpoints on a blindness-blurriness continuum. Each predicts that errors concentrate *only* in particular categories of games. The NM was directionally superior to the BM for explaining the observed error distribution, but participant errors are noisier than either archetype allows. Therefore, I simulated alternative *k*-level models that inject noise in different ways—by corrupting payoffs across final nodes (blurring) or within single final nodes (blinding)—to ask whether any of these models can more closely reproduce the pattern of errors across game groups.

To test whether generic noise at the deepest nodes can mimic the observed error pattern, I simulated Random Transpose Models (RTMs) and Random Default Models (RDMs). RTMs swap a pair of payoffs in $\{C_1, C_2, D_1, D_2\}$ at random. RDMs replace subsets of $\{C_1, C_2, D_1, D_2\}$ with values sampled from {1, 2, 3, 4} (mean = 2.5), treating 2.5 as a "default" value. I ran all 6 pairwise RTMs and all 16 RDMs and compared simulated and empirical error proportions [G1, G2, G3, G4] = [0.3094, 0.1747, 0.3094, 0.2066] using $fit = 1 - \sum_{g=1}^{4} |\widehat{g} - g|$, where $\widehat{g}$ is the observed error proportion and g is the simulated error proportion (1 = perfect match; −1 = worst case).

Randomizing all four payoffs produces the best match among these model versions, with [G1, G2, G3, G4] ≈ [0.295, 0.205, 0.295, 0.205] and fit ≈ 0.93 (see Table 3). For both the RDM and the RTM, fits are consistently better when the corruption spans both nodes for a given player than when it is confined to a single node across both players. While randomizing (or flipping) payoffs within a node is not the same as ablating the entire node, randomization produces a similar signature: randomizing node C distributes errors to game groups 3 and 4, as the ignore-C version of the BM predicts, and randomizing node D distributes errors to game groups 1 and 2, as the ignore-D version of the BM predicts. By contrast, distributing the deterioration across nodes (akin to NM blurriness) produces the same qualitative error signature as the NM—(G1 + G3) > (G2 + G4)—and fits the data far better.

While the models above place noise strictly at the deepest nodes, I tested a Noise Propagation Model (NPM) that introduces noise at the root node and amplifies it with distance from the root node. This idea echoes the Noisy Introspection Model (Goeree & Holt, 2004) where uncertainty about others' actions compounds with each reasoning level. The NPM expresses the existence of baseline noise even at the initial choice point but that grows with the computational complexity of representing ever more nodes and choices with each recursive step. The NPM computes SoftMax choice probabilities at each decision point using a base temperature τ_0 that doubles with depth (1.2 → 2.4 → 4.8) and is divided by k such that higher levels of Theory of Mind reasoning increase the resolution of the representation. The model uses backward induction, computing expected payoffs via these probabilities until it collapses the choice probabilities into a "stay"/"move" decision at the root node.

With $\tau_0 = 1.2$ and over 1,000,000 iterations per setting, the NPM's fit rises, peaks, and then falls as the probability of lapsing to 1st-order reasoning, $p(k = 1)$, varies. The best match occurs around $p(k = 1) \approx 0.5$ with error shares [0.308, 0.192, 0.308, 0.192] and a fit of ~0.959.

While exploring the full space of distance metrics and noise functions lies beyond the scope of this paper, the strength of this fit supports the plausibility of such models—ones that blur representations progressively with depth in the tree, rather than ablating nodes outright, and in doing so provide a more gradual and psychologically realistic account of strategic reasoning under pressure.

3.3. Order effects and controlling for the myopia-underestimation distinction

I analyzed error rates across rounds 1–8 (block 1), 9–16 (block 2), 17–24 (block 3), and 25–32 (block 4) to assess whether participants adjusted their reasoning over the course of the experiment. These errors were classified according to the criteria of the Near Sight Model.

The data revealed a slight decrease in error rates over time, suggesting a modest learning effect. The mean error rates decreased from 52.4 % in Block 1 to 43.9 % in Block 4. A non-parametric Friedman test indicated a statistically significant effect of trial order on error rates, $\chi^2(3) = 9.030$, $p = 0.029$, with a small effect size (Kendall's $W = 0.036$). Post-hoc pairwise comparisons using the Wilcoxon signed-rank test revealed that the only significant difference was between Block 1 and Block 4 ($p = 0.0068$). After applying a Bonferroni correction for six comparisons (adjusted $\alpha = 0.0083$), this difference remained statistically significant (adjusted $p = 0.0405$).

Errors in higher-order Theory of Mind tasks may indicate the chooser's cognitive limitations (myopia) or that the chooser believes their counterpart to be thinking at a lower level (underestimation). By facing participants against a consistently high-level opponent across all rounds, I provided participant the opportunity to learn that their counterpart was reasoning at a high level and provided no evidence to the contrary. Although one cannot conclusively rule out the possibility that participants underestimated their opponents, this slight learning effect makes that interpretation less likely. While underestimation may still affect results, it is more plausible that participants learned to respect the consistent high-level reasoning of their opponent and that the remaining errors can be attributed to cognitive limitations on the part of the participants.

Regardless, the myopia-underestimation ambiguity does not subtract from the differential error patterns observed across game types, which provides robust support for the Near Sight Model (NM) over the Blind Spot Model (BM) as a more accurate model of higher-order Theory of Mind under cognitive load.

4. Discussion

This study tested two competing models of higher-order Theory of Mind (ToM+) under cognitive load: the Blind Spot Model (BM) and the Near Sight Model (NM). Both assume that fully capable reasoners represent the entire game tree and solve it via backward induction, but they differ in how they handle depleted cognitive resources at the base case. The BM predicts "blindness," discarding distant possibilities, while the NM predicts "blurriness," approximating these distant states probabilistically. Although both models are plausible—BM aligns with the tendency to ignore complex information, and NM aligns with bounded rationality—our data strongly support the NM. Under time pressure, participants did not simply ignore distant outcomes but treated them as probabilistic averages. In contrast, BM's prediction of selective blindness was not borne out.

These findings strengthen the methodological basis for using sequential discrete-choice games to investigate higher-order Theory of Mind, as the diagnostic utility of such tasks depends critically on the psychological validity of the underlying model. Prior research employing these games (Flobbe et al., 2008; Goodie et al., 2012; Meijering et al., 2011; Meijering et al., 2012; Meijering et al., 2014; Zhang et al., 2012; Zhen & Yu, 2021) often relied on the "ignore-D heuristic"—the assumption central to the Blind Spot Model—to classify games as diagnostic. By identifying the Near Sight Model as more psychologically plausible, this study allows reinterpretation of previous findings and supports the design of future experiments with increased theoretical accuracy. Sequential-choice games informed by NM can be expanded to study richer ToM+ phenomena, such as incorporating hidden states to examine higher-order false beliefs or uncertainty. Indeed, Table 8 in the Appendix demonstrates how this framework can scale to third-order ToM. While complementary paradigms like the *p*-beauty contest have their own methodological merits, sequential discrete-choice games, refined with NM, now offer a tractable and interpretable approach for advancing research into recursive social reasoning.

More broadly, this research points toward a generalized cognitive framework for how people represent and reason through complex real-world social scenarios. Although my tasks involved structured games with common knowledge and explicitly presented payoffs, the cognitive principles uncovered here may extend beyond such controlled settings. Everyday social interactions, from interpreting sarcasm, irony, and gossip to managing deception, navigating trust, negotiating romantic relationships, and even strategic reasoning in nuclear deterrence, require imagining branching possibilities contingent on agentic decisions and uncertain outcomes. These mental representations are inherently recursive and tree-like, with each node encoding not only outcomes but also embedded mental states: beliefs, desires, intentions, misperceptions, and higher-order recursive beliefs about others. The Near Sight Model suggests that, under cognitive constraints, individuals do not discard distant possibilities entirely but rather represent them with gradually diminishing resolution, effectively treating distant agentic choices as probabilistic chance events. This gradual representational decay suggests a continuum of cognitive fidelity, extending from fully represented immediate nodes to increasingly blurred distant possibilities.

Real-world reasoning rarely provides explicit, externally structured decision trees. Instead, individuals must generate their own mental representations within constraints imposed by attention, working memory, and motivation. While NM currently implements discrete recursive steps, a more realistic representation may involve a continuous decay function, perhaps sigmoid-shaped, in which the clarity of node attributes gradually diminishes with increasing distance from the immediate decision point or as cognitive demands accumulate. These self-constructed mental trees are inevitably finite; they must terminate in leaf nodes representing vague expectations when cognitive resources reach their limits. Whether these terminal nodes result from never being constructed initially or from a Blind Spot-like pruning process, some version of selective omission remains psychologically plausible and indeed inevitable. Thus, while the present findings argue against BM as a comprehensive explanation, they suggest a nuanced reconciliation: selective omission and probabilistic blurring likely coexist as complementary strategies employed by finite cognitive systems attempting to effectively navigate through infinite social possibilities.

Nevertheless, several open questions and limitations should be acknowledged. In the version of the *Morality Game* run here, artificial agents responded instantly, rather than after a delay to appear human.

As mentioned before, participants only knew that their counterpart responded quicker than they did in the roughly half of games where they submitted a prediction, yet this speed may have been suspicious to some participants. This study focused exclusively on two-player sequential binary-choice games with fixed tree structures under conditions of time pressure as the sole form of cognitive load. Future research could test whether probabilistic heuristics similarly apply under alternative forms of load, such as multitasking or working memory constraints. While using a high-level robotic opponent helped reduce the ambiguity between cognitive myopia and opponent underestimation, future studies would benefit from explicitly measuring participants' beliefs about their counterparts' mental capacities and identities. Additionally, although minimal learning occurred within the short duration of this study, longer-term exposure to these tasks might shift participants' reliance between heuristics and explicit recursive reasoning. Eye-tracking data (Meijering et al., 2012) also indicate that participants often employ forward-search strategies rather than strict backward induction. Crucially, the NM's probabilistic blurring of distant states is compatible with forward reasoning as well: distant future states would be represented less clearly and probabilistically regardless of scan direction. Future research should explicitly examine how the Near Sight heuristic applies to forward-scanning cognitive strategies. Lastly, this experiment intentionally examined a simplified form of Theory of Mind, with participants operating under common knowledge, explicitly defined rules and states, and purely self-interested objectives. Real-world social cognition typically involves more complex mental-state attributions, including hidden beliefs, diverse desires, and context-dependent intentions. Thus, caution is warranted in generalizing my results to scenarios requiring rich inferential reasoning about varied mental contents. Nevertheless, understanding this foundational ToM ability provides a baseline for future investigations into how fully developed social cognition emerges from more basic recursive reasoning mechanisms.

Despite these caveats, the present study contributes a fundamental insight into higher-order social cognition: human reasoning about recursive mental states is neither perfectly rational nor purely myopic, but probabilistically approximate. This probabilistic near-sightedness likely reflects an adaptive cognitive strategy—one that achieves robust social reasoning despite resource constraints. Future research can now build upon these results to explore how this heuristic shapes diverse social cognitive phenomena, from communication and strategic deception to moral judgment. In doing so, researchers can better understand not only the boundaries of recursive reasoning but also how human cognition operates within those boundaries, approximating infinite social complexities within the finite mind.

CRediT authorship contribution statement

Gregory N. Stanley: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Ethical approval

This study was reviewed and granted an exemption from ongoing IRB review by the University of Michigan Institutional Review Board (IRB HSBS) under Exemption 3(i)(A) and/or 3(i)(B) at 45 CFR 46.104 (d). The exemption was issued on 11/22/2023 for the study titled "The Morality Game: Online Multiplayer Experiment Platform" (eResearch ID: HUM00243214). All participants provided prospective informed consent, and the research involved benign behavioral interventions with no identifiable information linked to the subjects.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used OpenAI's ChatGPT-4o and o1 in order to brainstorm and enhance readability. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This research was supported by the resources provided to the author as part of their Ph.D. program in the University of Michigan Psychology Department.

Declaration of competing interest

The author declares no conflicts of interest related to this study.

Acknowledgments

I would like to thank Dr. Jun Zhang and Dr. Rick Lewis for their invaluable suggestions and feedback on the conceptualization and design of this study and my other dissertation committee members Andras Molnar and Mu Zhang. I am grateful to the students in my Multiplayer Psychology of Morality and Cooperation Capstone course at the University of Michigan for their contributions to an earlier version of this study conducted as part of a class project: Nadine M. Bazzi, Brayden Noelle Moore, Bethany H. Tan, and Rongxuan Tian. Their experimentation and thoughtful discussions helped clarify key assumptions about the relationship between time pressure and cognitive load, which informed the present study. I also wish to thank the talented undergraduate research assistants who contributed to the development of the *Morality Game* platform: Alan Yang, Rohan Desai, Roujia Wang, Ekdatha Arramreddy, Liam Tsimhoni, Shulin Pan, Vijairam Moorthy, Michael Ahn, William Mathias, Mai Xu, James Kessler, Yahan Xing, Yang Li, and Duoming Bian. Their work in building this versatile online behavioral game theory platform has enabled a wide range of experiments, including the present study. I look forward to coauthoring a methods paper with them that will fully detail the platform's capabilities and their substantial contributions. I also thank the University of Michigan Psychology Department for their support. I appreciate the suggestions from the reviewers and the editor Neil Robert Bramley, which improved the paper greatly.

Appendix A

Pseudocode For Computing Expected Payoffs Via The Near Sight Model

```
FUNCTION expected_payoffs_near_sight_model(node, level_k)
  IF node.isLeaf() THEN
    RETURN node.payoffs as list

  Initialize parent_payoffs to node.payoffs
  IF level_k > 0 THEN
    // If cognitive resources are not yet exhausted
    Initialize child_payoffs as empty list
    // Recursively calculate payoffs for each child, reducing the ToM level
    FOR EACH child IN node.options DO
      Append expected_payoffs_near_sight_model(child, level_k - 1) to child_payoffs
    END FOR
    // Determine which child offers the best payoff for the chooser
    Set chooser_index to the index where node.chooser is TRUE
    Set chosen_payoffs to the maximum of child_payoffs using chooser_index
    // Combine the best child's payoffs with the parent's
    Update expected_payoffs by summing corresponding elements of parent_payoffs and chosen_payoffs
  ELSE
    // Base Case: Near Sight Heuristic: Compute expected values based on child probabilities
    Set expected_payoffs to parent_payoffs
    FOR EACH child IN node.options DO
      FOR i FROM 0 TO length(expected_payoffs) - 1 DO
        expected_payoffs[i] += expected_payoffs_near_sight_model(child, level_k - 1)[i] * child.probability
      END FOR
    END FOR
  END IF
  RETURN expected_payoffs
END FUNCTION
```

Pseudocode For Computing Expected Payoffs Via The Blind Spot Model

```
FUNCTION expected_payoffs_blind_spot_model(node, level_k)
  IF node.isLeaf() THEN
    RETURN node.payoffs as list

  Initialize parent_payoffs to node.payoffs
  IF level_k > 0 THEN
    // If cognitive resources are not yet exhausted
    Initialize child_payoffs as empty list
    // Recursively calculate payoffs for each child, reducing the ToM level
    FOR EACH child IN node.options DO
      Append expected_payoffs_blind_spot_model(child, level_k - 1) to child_payoffs
    END FOR
    // Determine which child offers the best payoff for the chooser
    Set chooser_index to the index where node.chooser is TRUE
    Set chosen_payoffs to the maximum of child_payoffs using chooser_index
    // Combine the best child's payoffs with the parent's
    Update expected_payoffs by summing corresponding elements of parent_payoffs and chosen_payoffs
  ELSE
    // Base Case: Blind Spot Heuristic: Blind to all but the first child's payoffs
    Set first_child_payoffs to expected_payoffs_blind_spot_model(node.options[0], level_k - 1)
    Update expected_payoffs by summing corresponding elements of parent_payoffs and first_child_payoffs
  END IF
  RETURN expected_payoffs
END FUNCTION
```

Fig. 7: Pseudocode for each k-level model: Both functions model recursive Theory of Mind reasoning at various cognitive depths (level_k). These models are implemented to calculate expected payoffs in a game theoretical context, using recursive depth-first traversal of a decision tree represented as nodes. The models operate identically in recursive descent until the cognitive resources (level_k) are depleted. At this exhaustion point, the base case logic differs significantly between models: the Near Sight Model computes expected values using probabilities of outcomes, whereas the Blind Spot Model simplifies the decision to the outcomes of the first child node only. This delineation is crucial for understanding the operational distinctions between how each model handles cognitive limitations in strategic decision-making scenarios. Readers are encouraged to note differences in the base case implementations for insights into each model's approach to cognitive resource constraints. See supplementary materials for Python code.

Table 4
Counterbalancing All 32 sequential-binary-choice game payoff structures: Payoffs are listed for each player: (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2). The "up-down pattern" summarizes whether a player's payoffs increase or decrease from the initial (e.g. A_1), through the intermediate (e.g. B_1), to the final node [e.g. max(C_1, D_1)]. Quadruplets compactly represent predicted choices—"move" (M) or "stay" (S)—for both models (BM and NM) at two cognitive depths ($k = 2$ and $k = 1$), in order (BM $k = 2$, BM $k = 1$, NM $k = 2$, NM $k = 1$). Starting payoffs ($A_1 = 2$ or 3), up-down patterns, and optimal initial move were systematically counterbalanced within and across game groups to prevent participants from relying on simple associations (e.g., "move if $A_1 = 2$") rather than genuine, higher-order reasoning. Diagnostic games are flagged (BM diagn/NM diagn) where a model's k = 2 and k = 1 choices diverge, and thus high error rates are expected. Game categories: (1) 1–8: both models predict frequent errors; (2) 9–16: BM only; (3) 17–24: NM only; (4) 25–32: neither.

game index	payoffs player 1	payoffs player 2	up-down pattern	BM $k = 2$	BM $k = 1$	NM $k = 2$	NM $k = 1$	quad- ruplet	BM diagn	NM diagn
1	(2, 1, 3, 4)	(1, 3, 4, 2)	↓↑-↑↓	stay	move	stay	move	SMSM	TRUE	TRUE
2	(2, 1, 3, 4)	(3, 2, 4, 1)	↓↑-↑↓	stay	move	stay	move	SMSM	TRUE	TRUE
3	(2, 1, 3, 4)	(4, 2, 3, 1)	↓↑-↑↓	stay	move	stay	move	SMSM	TRUE	TRUE
4	(3, 4, 1, 2)	(2, 3, 1, 4)	↑↓-↓↑	stay	move	stay	move	SMSM	TRUE	TRUE
5	(2, 1, 3, 4)	(2, 3, 1, 4)	↓↑-↓↑	move	stay	move	stay	MSMS	TRUE	TRUE
6	(3, 1, 2, 4)	(2, 3, 1, 4)	↓↑-↓↑	move	stay	move	stay	MSMS	TRUE	TRUE
7	(3, 2, 1, 4)	(3, 2, 1, 4)	↓↑-↓↑	move	stay	move	stay	MSMS	TRUE	TRUE
8	(3, 4, 1, 2)	(3, 2, 4, 1)	↑↓-↑↓	move	stay	move	stay	MSMS	TRUE	TRUE
9	(2, 1, 3, 4)	(2, 3, 4, 1)	↓↑-↑↓	stay	move	stay	stay	SMSS	TRUE	FALSE
10	(3, 4, 1, 2)	(1, 3, 2, 4)	↑↓-↓↑	stay	move	stay	stay	SMSS	TRUE	FALSE
11	(3, 4, 1, 2)	(3, 2, 1, 4)	↑↓-↓↑	stay	move	stay	stay	SMSS	TRUE	FALSE
12	(3, 4, 1, 2)	(4, 2, 1, 3)	↑↓-↓↑	stay	move	stay	stay	SMSS	TRUE	FALSE
13	(2, 1, 3, 4)	(3, 2, 1, 4)	↓↑-↓↑	move	stay	move	move	MSMM	TRUE	FALSE
14	(2, 3, 1, 4)	(3, 2, 4, 1)	↑↑-↑↓	move	stay	move	move	MSMM	TRUE	FALSE
15	(2, 4, 1, 3)	(2, 3, 4, 1)	↑↓-↑↓	move	stay	move	move	MSMM	TRUE	FALSE
16	(3, 4, 1, 2)	(2, 3, 4, 1)	↑↓-↑↓	move	stay	move	move	MSMM	TRUE	FALSE
17	(2, 1, 4, 3)	(1, 3, 2, 4)	↓↑-↓↑	stay	stay	stay	move	SSSM	FALSE	TRUE
18	(2, 1, 4, 3)	(3, 2, 1, 4)	↓↑-↓↑	stay	stay	stay	move	SSSM	FALSE	TRUE
19	(2, 1, 4, 3)	(4, 2, 1, 3)	↓↑-↓↑	stay	stay	stay	move	SSSM	FALSE	TRUE
20	(3, 4, 2, 1)	(2, 3, 4, 1)	↑↓-↑↓	stay	stay	stay	move	SSSM	FALSE	TRUE
21	(2, 1, 4, 3)	(2, 3, 4, 1)	↓↑-↑↓	move	move	move	stay	MMMS	FALSE	TRUE
22	(3, 1, 4, 2)	(2, 3, 4, 1)	↓↑-↑↓	move	move	move	stay	MMMS	FALSE	TRUE
23	(3, 2, 4, 1)	(3, 2, 4, 1)	↓↑-↑↓	move	move	move	stay	MMMS	FALSE	TRUE
24	(3, 4, 2, 1)	(3, 2, 1, 4)	↑↓-↓↑	move	move	move	stay	MMMS	FALSE	TRUE
25	(2, 1, 4, 3)	(2, 3, 1, 4)	↓↑-↓↑	stay	stay	stay	stay	SSSS	FALSE	FALSE
26	(3, 4, 2, 1)	(1, 3, 4, 2)	↑↓-↑↓	stay	stay	stay	stay	SSSS	FALSE	FALSE
27	(3, 4, 2, 1)	(3, 2, 4, 1)	↑↓-↑↓	stay	stay	stay	stay	SSSS	FALSE	FALSE
28	(3, 4, 2, 1)	(4, 2, 3, 1)	↑↓-↑↓	stay	stay	stay	stay	SSSS	FALSE	FALSE
29	(2, 1, 4, 3)	(3, 2, 4, 1)	↓↑-↑↓	move	move	move	move	MMMM	FALSE	FALSE
30	(2, 3, 4, 1)	(3, 2, 1, 4)	↑↑-↓↑	move	move	move	move	MMMM	FALSE	FALSE
31	(2, 4, 3, 1)	(2, 3, 1, 4)	↑↓-↓↑	move	move	move	move	MMMM	FALSE	FALSE
32	(3, 4, 2, 1)	(2, 3, 1, 4)	↑↓-↓↑	move	move	move	move	MMMM	FALSE	FALSE

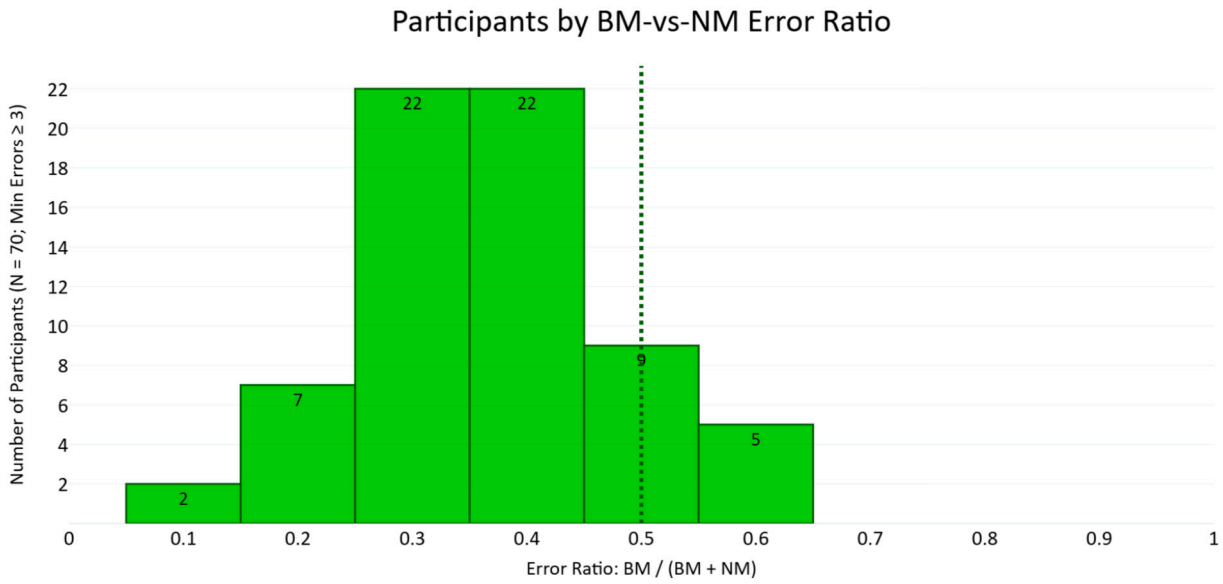


Fig. 8: Distribution of participants by BM vs NM error ratio: To determine if the BM described a subpopulation of participants despite the general trend, I scored participants by the ratio of their errors that were expected by each model. Only the BM expects errors in game group 2, while only the NM expects errors in game group 3, and so these ratios are the number of errors in game group 2 divided by the total number of errors in game groups 2 and 3. This excludes participants who only made 2 or fewer errors in game groups 2 and 3. 53 participants made errors more consistent with the NM, 6 participants made errors more consistent with the BM, and 9 participants had error ratios of 0.5, meaning a tie. Of the 5 participants with error ratios greater than 0.5, all had error ratios less than 0.7. This does not provide strong support for the notion that the BM describes a significant subpopulation of participants.

Table 5
Number of participants by number of abdications (Non-responses):

N Abdications	0	1	2	3	4	5	6	7	8	10	14	32
N Participants	38	12	10	6	1	1	1	1	1	2	1	1

Table 6
Error rates by game group with and without abdications counted as errors: In the primary analysis, abdications (timeouts) were excluded because they represent non-responses rather than strategic choices and thus cannot be mapped onto either the Blind-Spot or Near-Sight models. For completeness, I repeated the analysis treating abdications as if they were errors. This robustness check shows that the main conclusions are unchanged. When abdications are counted as errors, the key contrast remains highly significant: (Groups 1 & 3 vs. 2 & 4): $X^2(1, N = 2336) = 103.04, p = 3.3 \times 10^{-24}$, *Cramér's V* $= 0.149$.

BM Expects More Errors	NM Expects More Errors	Error Rate Without Abdications	Error Rate With Abdications	Δ Rate
TRUE	TRUE	0.548	0.562	+0.014
TRUE	FALSE	0.315	0.349	+0.034
FALSE	TRUE	0.570	0.599	+0.029
FALSE	FALSE	0.368	0.390	+0.023

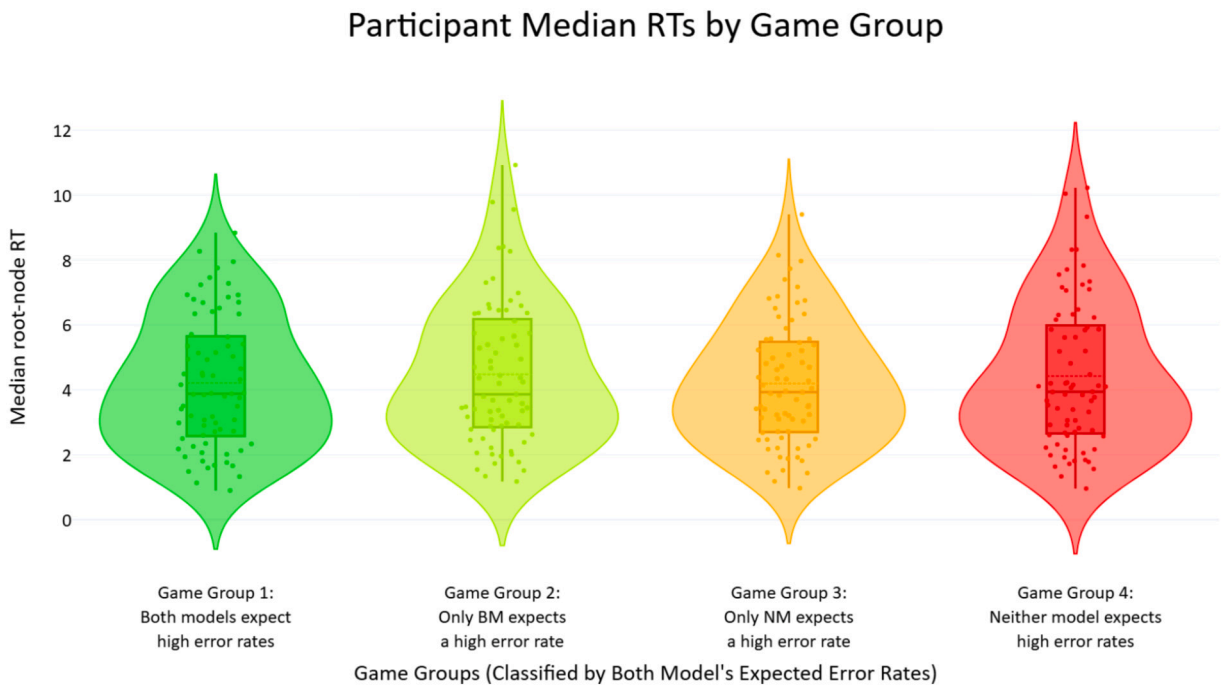


Fig. 9. Median root node reaction times by game group: Each dot is a participant's median RT for non-abdicated root choices; violins show the distribution and boxes show the interquartile range. Game groups: G1 = both models expect high error; G2 = only BM expects high error; G3 = only NM expects high error; G4 = neither. Across groups, median RTs did not significantly differ (Friedman $\chi^2(3) = 6.58$, $p = 0.086$; $N = 73$).

Table 7
Choice-prediction consistency: 77.6 % of predictions were made at the same ToM level as the preceding choice, indicating that participants maintained a coherent ToM level within each game. Trials were coded as *coherent* when the Theory-of-Mind (ToM) level implied by the root action matched the ToM level implied by the prediction (CC for correct–correct; MM for mistaken–mistaken), and *incoherent* otherwise (CM or MC). In move-leading quadruplets (MSMS, MSMM, MMMS, MMMM), move (M) is the optimal (correct) higher-order root node choice, while in stay-leading quadruplets (SMSM, SMSS, SSSM, SSSS), stay (S) is the optimal (correct) higher-order root node choice. Because predictions are only solicited after a *move* at the root, the move-leading quadruplets sample predictions after root actions that are *correct*, whereas the stay-leading quadruplets sample predictions after *mistaken* root actions. Predictions were coded as *correct* when they expected player 2 to choose at the 1st-order and *mistaken* when they expected player 2 to choose at the 0th-order.

Choice-Prediction Consistency in Move-leading Quadruplets			
Quadruplet	Correct-Correct/*N*	Correct-Mistaken/*N*	*N* (total)
MSMS	0.840	0.160	156
MSMM	0.729	0.271	229
MMMS	0.804	0.196	143
MMMM	0.784	0.216	218

Choice-Prediction Consistency in Stay-leading Quadruplets			
Quadruplet	Mistaken-Mistaken/*N*	Mistaken-Correct/*N*	*N* (total)
SMSM	0.791	0.209	182
SMSS	0.750	0.250	124
SSSM	0.762	0.238	181
SSSS	0.757	0.243	140

Table 8
Future experiment - diagnosing 3rd-order theory of mind through sequential games.

Panel A: Optimal Moves for Different Levels in Each Game			
Game Label	3rd-Order Optimal Choice	2nd-Order Optimal Choice	1st-Order Optimal Choice
SMM?	Stay (S)	Move (M)	Move (M)
MSM?	Move (M)	Stay (S)	Move (M)
SSM?	Stay (S)	Stay (S)	Move (M)

Panel B: Diagnosing ToM Level from Response Patterns Across Games	
Response Pattern Across Games (Order: SMM, MSM, SSM)	Corresponding ToM Level Diagnosis
MMM	1st-Order
MMS	No Clear Diagnosis
MSM	1st- or 2nd-Order
MSS	2nd-Order
SSS	2nd- or 3rd-Order
SSM	No Clear Diagnosis
SMS	3rd-Order
SMM	1st- or 3rd-Order

Panel A represents the optimal root-node choice for each game (SMM?, MSM?, SSM?) across 3rd-, 2nd- and 1st-order players. The choice pattern (e.g., SMM) indicates the optimal move for each reasoning level: "Stay" or "Move" as determined by the players' perspective. This serves as the foundation for understanding how each level of Theory of Mind is diagnosed based on participants' root-node choices.
Panel B demonstrates how participants' response patterns across a set of three games—SMM, MSM, SSM—can be used to diagnose their Theory of Mind level. The response pattern (e.g., "MMM" or "SMS") refers to the participant's choices across three games in a specific order. For each pattern:
The ToM level diagnosis is based on identifying which level's predicted choices are consistent across all three games.
For example, a response pattern of "SMS" indicates that the participant stayed in Game SMM, moved in Game MSM, and stayed in Game SSM. This aligns with the predicted 3rd-order choices across these games, resulting in a 3rd-order diagnosis.
Game Notation (Panel A): SMM, MSM, SSM denote the game label—the optimal move sequence expected for players with different levels of reasoning.
Response Pattern (Panel B): The patterns like MMM or SMS represent the choices a participant made across the three games.
This table illustrates the methodology for diagnosing participants' Theory of Mind (ToM) level up to the 3rd-order using a set of alternating-choice games. The diagnosis is made based on the root-node choice that participants make for each game. The example games—SMM, MSM, and SSM—each offer unique payoff structures that elicit different optimal choices for players with 3rd-, 2nd-, and 1st-order reasoning.
Panel A shows the predicted optimal moves for each game across players with different levels of ToM.
Panel B summarizes the possible response patterns across the three games and their corresponding ToM level diagnosis. The response pattern indicates the choices participants made across three trials, and each pattern is evaluated to determine which level of recursive reasoning is consistent across all games. For example, the response pattern "MMM" aligns with the choices expected for a 1st-order player, resulting in a clear 1st-order diagnosis.

This structured approach enables researchers to categorize participants' reasoning capabilities systematically, and extending this method to multiple trials increases confidence in the diagnosis of a given level of Theory of Mind. The presented methodology serves as an example of how higher-order ToM can be diagnosed in a novel, precise manner, paving the way for new and more complex studies into human social cognition.

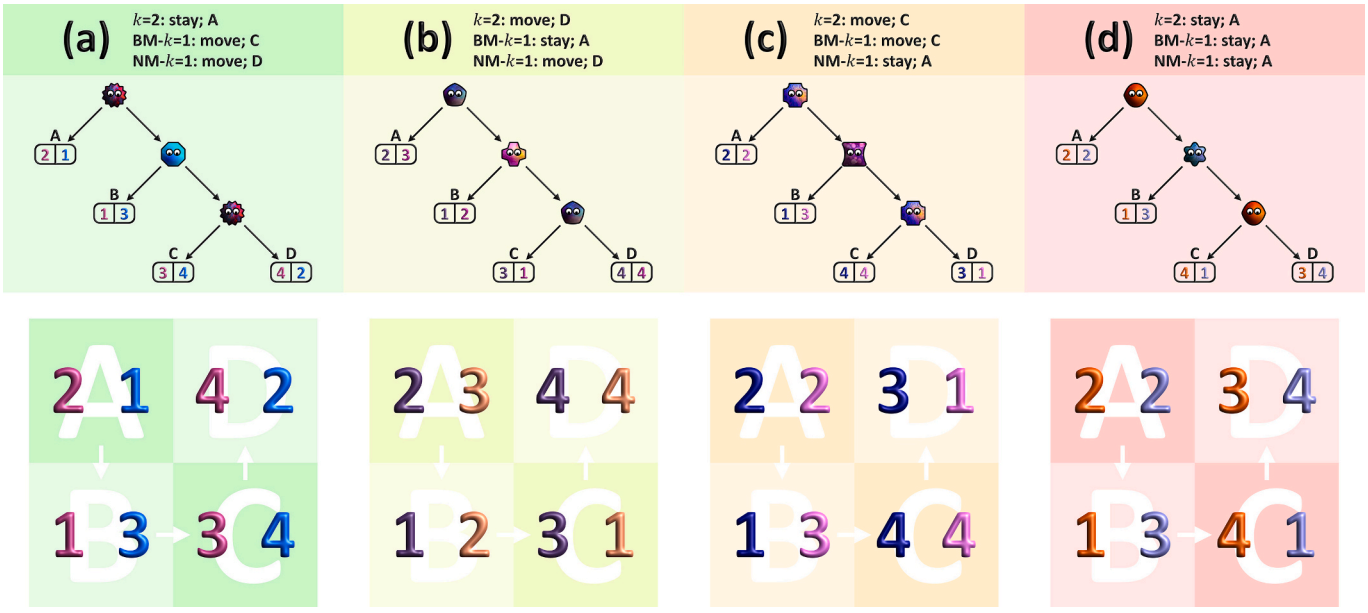


Fig. 10. Spatial versus non-spatial game representations: the top row depicts four games, mimicking the non-spatial game tree stimuli of the *Morality Game*. The bottom row shows the same games, mimicking the spatial game boards from Hedden and Zhang (2002). Players start in cell A, progress downward to cell B, rightward to cell C, and upward to cell D. This spatial game representation explains why Hedden and Zhang may have created an 'ignore-D heuristic', instead of an 'ignore-C heuristic', even though the choice of node to ignore becomes arbitrary in a game tree representation, which is non-spatial in the sense that sibling node order is strategically irrelevant.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106331.

Data availability

The data and analysis script associated with this study are openly available on the Open Science Framework at https://doi.org/10.17605/OSF.IO/M8XEJ. The *Morality Game* experiment platform can be found at https://moralitygame.com.

References

Alaoui, L., & Penta, A. (2016). Endogenous depth of reasoning. *The Review of Economic Studies, 83*(4), 1297–1333.

Arslan, B., Taatgen, N. A., & Verbrugge, R. (2017). Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: a computational modeling study. *Frontiers in Psychology, 8*, 275.

Baron, J., & Ritov, L. (1992). Status-quo and omission biases. *Journal of Risk and Uncertainty, 5*(1), 49–61.

Bernheim, B. D. (1984). Rationalizable Strategic Behavior. *Econometrica, 52*(4), 1007–1028.

Broadbent, D. E. (1958). *Perception and communication*. New York: Pergamon Press.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., … Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games, 4*(1), 1–43.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 861–898.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215.

De Weerd, H., Verbrugge, R., & Verheij, B. (2015). Higher-order theory of mind in the tacit communication game. *Biologically Inspired Cognitive Architectures, 11*, 10–21.

De Weerd, H., Verbrugge, R., & Verheij, B. (2022). Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems, 36*(2), 30.

Doshi, P., & Gmytrasiewicz, P. J. (2009). Monte Carlo sampling methods for approximating interactive POMDPs. *Journal of Artificial Intelligence Research, 34*, 297–337.

Estes, D., Wellman, H. M., & Woolley, J. D. (1989). Children's understanding of mental phenomena. *Advances in Child Development and Behavior, 22*, 41–87.

Flavell, J. H. (1977). *Cognitive Development*. Englewood Cliffs, NJ: Prentice-Hall.

Flavell, J. H., Botkin, P., Fry, C. W., & Jarvis, D. (1968). *The Development of Role-Taking and Communication Skills in Children*. New York: Wiley.

Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information, 17*, 417–442.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349* (6245), 273–278.

Goeree, J. K., & Holt, C. A. (2004). A model of noisy introspection. *Games and Economic Behavior, 46*, 365–382.

Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of Theory-of-Mind Reasoning in Competitive Games. *Journal of Behavioral Decision Making, 25*, 95–108.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, 145–171.

Hedden, T., & Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition*, 1–36.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology, 8*(3), Article e1002410.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence, 101*(1–2), 99–134.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.

Knuth, D. E., & Moore, R. W. (1975). An analysis of alpha-beta pruning. *Artificial Intelligence, 6*(4), 293–326.

Kocsis, L., & Szepesvári, C. (2006). Bandit based monte-carlo planning. In *European Conference on Machine Learning* (pp. 282–293). Berlin: Springer Berlin Heidelberg.

Krueger, J. I., & Acevedo, M. (2005). Social Projection and the Psychology of Choice. In D. A. M. D (Ed.), *The Self in Social Judgment* (pp. 17–41). New York: Psychology Press.

Krueger, J. I., DiDonato, T. E., & Freestone, D. (2012). Social projection can solve social dilemmas. *Psychological Inquiry, 23*, 1–27.

Land, A. H., & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometric, 28*(3), 497–520.
Laplace, P. S. (1814). *Essai Philosophique sur Les Probabilités*. Paris: Unknown.
Lieder, F., & Griffith, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, Article e1.
McNamara, R. S. (1962, June 16). *The no-cities doctrine*. Ann Arbor, Michigan, USA: University of Michigan, Ann Arbor.
Meijering, B., Rijn, H. V., Taatgen, N., & Verbrugge, R. (2011). I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In , *Vol. 33, No. 33. Proceedings of the Annual Meeting of the Cognitive Science Society*.
Meijering, B., Van Rijn, H., Taatgen, N. A., & Verbrugge, R. (2012). *What Eye Movements Can Tell About Theory of Mind in a Strategic Game*.
Meijering, B., Taatgen, N. A., Rijn, H. V., & Verbrugge, R. (2014). Modeling inference of mental states: As simple as possible, as complex as necessary. *Interaction Studies, 15*(3), 455–477.
Nagel, R. (1995). Unraveling in guessing games: an experimental study. *The American Economic Review, 85*(5), 1313–1326.
Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences, 36*(1), 48–49.
Nash, J. (1951). Non-cooperative games. *Annals of Mathematics, 54*(2), 286–295.
von Neumann, J., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton, N.J.: Princeton University Press.
Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*(3), 151.
Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.
Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge, MA: Cambridge University Press.
Pearce, D. G. (1984). Rationalizable behavior and the problem of perfection. *Econometrica, 52*(4), 1029–1050.
Pearl, J. (1984). *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc.
Perner, J. (1979). Young children's preoccupation with their own payoffs. *Developmental Psychology, 15*(2), 204–213.
Perner, J. (1994). Understanding the representational mind. *Pedagogische Studiën, 71*(3).
Perner, J., & Wimmer, H. (1985). "John Thinks That Mary Thinks That …" Attribution of Second-Order Beliefs by 5- to 1 O-Year-Old Children. *Journal of Experimental Child Psychology, 39*, 437–471.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.
Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence, 49*(1–3), 361–395.
Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson Education Inc.
Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.
Simon, H. A. (1957). *Models of Man: Social and Rational*. Hoboken, NJ: John Wiley & Sons.
Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization, 25*, 309–327.
Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: theory and experimental evidence. *Games and Economic Behavior, 10*(1), 218–254.
Waltz, K. (1981). The spread of nuclear weapons: More may be better. *Adelphi Papers, 21*(171), 1–32.
Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598–604.
Wellman, H. M., Carey, S., Gleitman, L., Newport, E. L., & Spelke, E. S. (1990). *The Child's Theory of Mind*. Cambridge, MA: MIT Press.
Zhang, J., Hedden, T., & Chia, A. (2012). Perspective-taking and depth of theory-of-mind reasoning in sequential-move games. *Cognitive Science, 36*(3), 560–573.
Zhen, S., & Yu, R. (2021). Neural correlates of recursive thinking during interpersonal strategic interactions. *Human Brain Mapping, 42*(7), 2128–2146.