Opponent-Aware Intuitive Gamer: Bayesian Inference of Sensibility in Novel Games

Michal Lewkowicz (michal01@mit.edu)
MIT

Aryan Naveen (aryannav@mit.edu)
MIT

Abstract

To gather insight into how humans reason in novel game settings, it is necessary to construct computational models of cognition that approximate human performance in resource constrained settings. In this work we seek to gain insight into how human's use fast heuristics to approximate action sensibility and simultaneously reason about an opponent's strategy to inform their own action selection. We use the Intuitive Gamer model as a proxy for a novice human player and extend it to incorporate probabilistic opponent modeling. This provides a more realistic framework for how human's dynamically adjust their heuristics based on presumed opponent competency. We present a mathematical formulation for *sensibility* which we leverage to characterize contexts where different policies diverge. Subsequently, we present an Opponent Aware Intuitive Gamer framework which leverages a Bayesian model to update the posterior over opponent policies and dynamically adjust its value function weights. We finally conduct several experiments to evaluate the difference between the baseline Intuitive Gamer, our extension, and a computationally exhaustive expert policy. Code is made publicly available at https://github.com/MLewkowicz/opponent-aware-intuitive-gamer.

Keywords: Bayesian, Computational Cognitive Science, Game Theory

Introduction

Games are an essential tool for understanding humans' reasoning abilities. Competitive two-player zero-sum games are a useful testbed for understanding how human's plan and act in resource constrained environments. Humans have an inherent computational constraint, as they cannot run exhaustive searches to evaluate end-game outcomes. Various works in both psychology and computer science have shown that humans create lightweight models for understanding structure in games, and extract compact value functions from these models to approximate their own advantage or disadvantage, without requiring extensive search across game states (Collins et al., 2025). These value functions are subsequently leveraged for probabilistically sampling a reasonable action.

Another useful feature of two-player zero-sum games is that they necessitate opponent modeling. The features of an action that a human player considers when determining if an action is sensible cannot be exclusively derived from the rule specifications of a game, but requires an estimate of an opponent's underlying policy and skill level. We frame this problem within a Bayesian modeling context. A human player must maintain a belief over the opponent's underlying policy or value function, and inform their own actions to maximize their own advantage and minimize their losses.

In this project, we build upon the work in Intuitive Gamer (Collins et al., 2025), where the authors construct a flat, goal-directed value function that can be quickly computed and probabilistically sampled to serve as a model for human judgment in novel game settings. However, the Intuitive Gamer framework only evaluates board positions one state into the future which does not factor in the opponent strategy and skill level. We extend this model to maintain a belief over the opponent's policy conditional on their action history with the assumption that the opponent can only be one of three predetermined policies. To achieve this, we will first formalize the notion of sensibility. We outline two metrics for quantifying how likely an action history is produced by a certain policy, and use these metrics to update our belief over the opponent.

Additionally, we propose that the Intuitive Gamer model needs to dynamically adjust its value function in response to this opponent policy estimate, or if the model recognizes the opponent is too advanced, modify its evaluation depth. We can imagine a setting where we are competing with a player who seems to be a novice in a game. This information would cause us to prescribe less attention to defensive strategies and more attention to fast offensive play. A classic example would be going for a four move checkmate against a chess novice. On the other hand, when competing against someone who is well versed in the game, more conservative and cautious game play is necessary.

Our main contributions for this work are as follows:

1. We define two metrics for quantifying move sensibility
2. We characterize states where policies diverge in their action likelihoods
3. We implement a Bayesian framework for modeling the opponent's policy
4. We extend the Intuitive Gamer model to dynamically adjust its value function weights and/or evaluation depth based on this belief distribution

Related Works/Background

Within the field of computer science, the study of games has largely been guided by an interest in *optimality*. Foundational game theorems such as Nash Equilibrium (Nash, 2024) attempt to characterize convergence states under rational play. While such research has culminated in amazing advances in computational game engines such as AlphaGo and Deep Blue

(D. Silver et al., 2016), these works are interested in matching or exceeding expert-level play that is built on countless hours spent studying the game mechanics. Most individuals, however, are not game experts but still posses the ability to quickly construct heuristics that guide them towards more advantageous positions.

This reality has motivated significant research in the computational cognitive science community in pursuit of capturing human cognition. Extensive work has leveraged foundational Bayesian principles to evaluate human ability to transfer prior knowledge when adapting to new domains. *Virtual Tools* (Allen, Smith, & Tenenbaum, 2020) explores this phenomenon by capturing how humans integrate structured priors on candidate tools and actions to compose plans for unseen task situations and continuously adapt these priors through simulations. In a similar vein, the work conducted in (T. Silver, Allen, Lew, Kaelbling, & Tenenbaum, 2019) explores how strategies can be learned from demonstrations to construct priors that generalize beyond the demonstrations through a probabilistic grammar prior that biased towards simple policies, in testament to Occam's Razor. Several other works (Allen, Smith, Piterbarg, Chen, & Tenenbaum, 2020) similarly argue that it is important to deriving flexible priors that allow for flexible transfer.

These works have demonstrated remarkable progress towards capturing human intelligence. However, in the context of games, an important concept is inference over future rewards and costs. Intuitive Gamer presents a novel approach to capture human's generalizable rationale for novel games through a light-weight goal directed planning framework (Collins et al., 2025).

Intuitive Gamer Overview Several works in both robotics and cognitive science explore leveraging game state evaluation functions and roll outs over decision trees to construct a policy $\pi(a_t \mid s_t)$ (Browne et al., 2012). However, Intuitive Gamer contends that such intense simulation and highly tuned value functions is not realistically employed by a novice player. Consequently, the authors propose a framework that leverages a more general-purpose abstract value function and a much shallower look ahead to stochastically select the action. Formally, the abstract value function is composed of three underlying utilities: (1) immediate progress towards player's own goal $U_{\texttt{self}}$, (2) progress blocking of opponent's goal $U_{\texttt{opp}}$, and (3) an auxiliary value that guides values of actions $U_{\texttt{aux}}$ (e.g: playing closer to the center).

$$\mathcal{V}(s_t, a_t) = U_{\texttt{self}}(s_t, a_t) + U_{\texttt{opp}}(s_t, a_t) + U_{\texttt{aux}}(s_t, a_t) \quad (1)$$

The Intuitive Gamer policy π is constructed at each state by computing the softmax of the value approximations according to the heuristic above. While Intuitive Gamer presents exciting progress in modeling a lightweight and transferable abstract heuristic, it does not consider the influence that opponent's actions have on shaping the abstract heuristic.

Opponent Modeling Extensive research has been conducted on trying to design frameworks that apply a Bayesian lens to reasoning about one's opponent. Work such as (Jara-Ettinger, Baker, Ullman, & Tenenbaum, 2025) attempts to infer the other agent's goal that best explains their past sequence of decisions. This is important because recursive agent models are present throughout daily life where one agent's utilities and motives are coupled with another's. Work such as (Zhi-Xuan, Mann, Silver, Tenenbaum, & Mansinghka, 2020) try to perform efficient Bayesian inference over an agent's goals and internal plan in order to estimate underlying intent from optimal and sub-optimal action trajectories.

Problem Formulation

As presented in Intuitive Gamer, we can mathematically formulate a game G as a composition of feasible states $\mathcal{S}$; possible actions $\mathcal{A}$; rule $\mathcal{T}$ specifying state transitions and valid actions; and goal functions mapping from states $r : \mathcal{S} \to \mathcal{R}$. We are interested in modeling a policy $\pi_G(a_t \mid s_t)$ for how to play without prior game play experience. We hope to extend the work of Intuitive Gamer by additionally considering the opponent's past actions to infer their underlying policy and plan accordingly $\pi_G(a_t \mid s_t, \pi_{\texttt{opp}})$.

In order to both infer and plan appropriately for our opponents skill level and underlying policy, we must formally define the concept of an *opponent posterior* and a move *sensibility*.

Definition (Opponent Posterior)**.** Let $\tau = \{(s_t, a_t)|$ if $t \,\%\, 2 = 1\}_{t=1}^{T}$ be the opponent state–action trajectory and let π denote a candidate policy. We define the Bayesian posterior probability of the opponent using policy π given trajectory τ as $P(\pi \mid \tau)$. This essentially quantifies how likely it is that trajectory τ was generated by policy π.

Definition (Sensibility)**.** We define the sensibility of a trajectory τ to be the likelihood that τ was generated given that policy π. We will use the notation $\text{Sens}(\tau \mid \pi)$ to denote this likelihood.

Now, in order to compute $P(\pi \mid \tau)$, we leverage the following Bayesian framework:

$$P(\pi \mid \tau) = \frac{\text{Sens}(\tau \mid \pi) P(\pi)}{P(\tau)}$$
$$P(\pi \mid \tau) \propto \text{Sens}(\tau \mid \pi) P(\pi)$$

Given the invariance of $P(\tau)$ to π, we can drop that term from the above expression. Additionally, for the prior over policies we assume a uniform distribution over candidates. However, this formulation allows for differing human priors over opponent competency.

In order to calculate $\text{Sens}(\tau \mid \pi)$ we present two distinct formulations, each of which captures a unique component of π decision making process.

1. **Action Likelihoods:** Given the stochastic definition of $\pi(a \mid s)$, we can directly compute $\text{Sens}(\tau \mid \pi)$ as follows:

$$\text{Sens}(\tau \mid \pi) = \prod_{(s,a)\in\tau} \pi(a \mid s)$$

2. **Agreement Likelihood:** We define the agreement likelihood as the fraction of state–action pairs in τ for which the observed action matches the most likely action under policy π:

$$\text{Sens}(\tau \mid \pi) = \frac{1}{|\tau|} \sum_{(s,a)\in\tau} \mathbb{I}\left[a = \arg\max_{a'} \pi(a' \mid s)\right] \quad (2)$$

where $\mathbb{I}[\cdot]$ is the indicator function.

To obtain a normalized distribution over candidate policies $\pi \in \Pi$, we apply a softmax over the policy likelihood:

$$P(\pi \mid \tau) = \frac{\exp(P(\pi \mid \tau))}{\sum_{\pi'\in\Pi} \exp(P(\pi' \mid \tau))} \quad (3)$$

Implementation Details

In order to experimentally verify our opponent modeling framework, we implement three policies that the opponent can be following in simulated play.

The first *opponent* policy is a **Random Policy** which uniformly samples valid actions regardless of the state features.

The second *opponent* policy is our re-implementation of **Intuitive Gamer** which computes the heuristic function for game states one action ahead, and probabilistically samples valid actions proportional to their softmax. For the opponent, the Intuitive Gamer model has weights 1 for each of the three components in its value function (1), as it most closely fits data from novice play according to experiments in (Collins et al., 2025).

Our third and final *opponent* policy is **Monte Carlo Tree Search (MCTS)**, which we use as a computational-intensive baseline for expert-level play. Given some iteration count, MCTS expands a game tree using the Upper Confidence Bound (UCB) heuristic to balance exploration of unvisited states and the exploitation of states that return high reward. At leaf nodes, the algorithm simulates random play until the game reaches a terminal state, receiving reward depending on the outcome (win, draw, loss). The rewards are then back-propagated up the tree, updating value estimates relative to the specific player acting at each node.

For the model we are pitting against the opponent policies, we extend the implementation of Intuitive Gamer to create **Opponent-Aware Intuitive Gamer**, which dynamically adjusts its weights given a belief distribution over opponent policies, according to our method described in (4). In simulated play, we recognize that even with opponent modeling and dynamic weight adjustment, it would be infeasible for our base Intuitive Gamer model that only does one-step look ahead to have a reasonable chance when versing an MCTS agent that has accumulated experience about which actions lead to desirable game outcomes. Therefore, we also implement our **Opponent-Aware Intuitive Gamer** with a depth parameter that allows for evaluating the weighted value function at some pre-specified horizon. Since the model does not have access to the sampling mechanism of the opponent policy, the model rolls out the top-k most reasonable actions according to its own policy up to some horizon, and evaluates the value function at the leaf nodes, where k is arbitrarily chosen to limit the size of the game search tree. The likelihood of the next actions is then proportional to the average of the value function across leaf nodes.

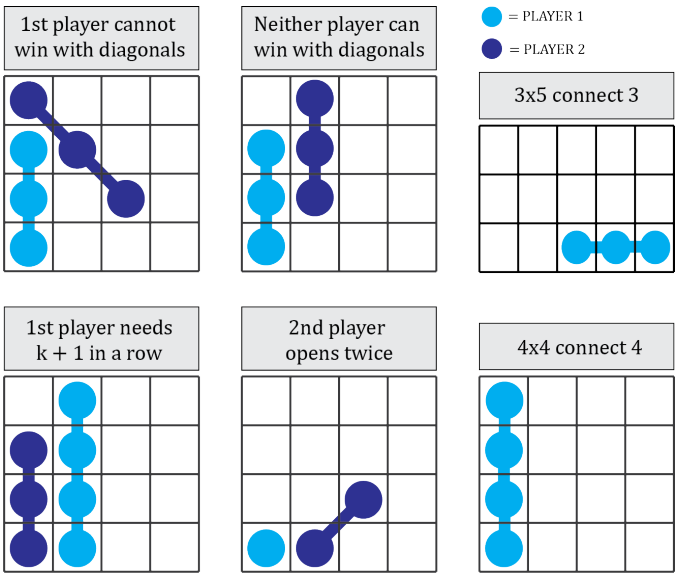


Figure 1: We visualize the game variants of connect-k games in $m \times n$ sized grids that we use in our analysis. We apply certain constraints to players, such as one player or neither player being able to win on diagonals, opening move asymmetries, and $k+1$ chains required for the 1st player.

Sensibility Experiments

In order to evaluate the accuracy of our sensibility definition, we conduct preliminary analysis to compare states where policies converge on similar actions and where they diverge across different game variants. To generate our game variants, we consider a subset of the 121 variants from Intuitive Gamer, which are all connect-k games in $m \times n$ sized grids. We induce more variation in our game generation by applying asymmetric rules to each of the players. We only consider three types of constraints: (1) either or both players cannot win with contiguous diagonal chains; (2) 2nd player can play two moves on their first turn; and (3) 1st player needs to connect $k+1$ in a row, while the 2nd player only needs k in a row. We visualize all the variants we consider in 1.

We formulate two metrics to evaluate the *agreement* between two policies for a given state, inspired by our definition of sensibility presented earlier:

1. **Max Action Agreement:** Given a game state $s \in \mathcal{S}$,

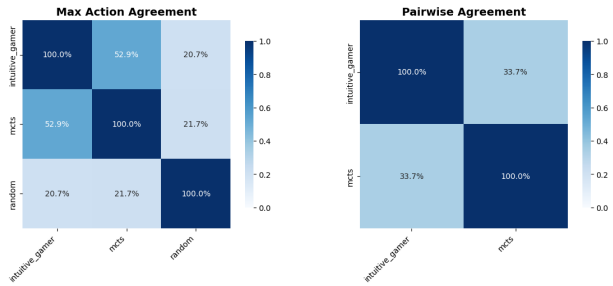


Figure 2: Consistency of action preferences across candidate policies evaluated over states in $\mathcal{S}$.

π_1 and π_2 are in agreement if $\arg\max_{a\in\mathcal{A}} \pi_1(a \mid s) = \arg\max_{a\in\mathcal{A}} \pi_2(a \mid s)$.

2. **Pairwise Comparison Agreement:** Given a game state $s \in \mathcal{S}$, we approximate the KL divergence between $\pi_1(a \mid s)$ and $\pi_2(a \mid s)$ through a sequence of pairwise comparisons, where we randomly sample 2 actions $a_1, a_2 \in \mathcal{A}$ and say the polices are in agreement if $(\pi_1(a_1 \mid s) - \pi_1(a_2 \mid s)) \cdot (\pi_2(a_1 \mid s) - \pi_2(a_2 \mid s)) \geq 0$.

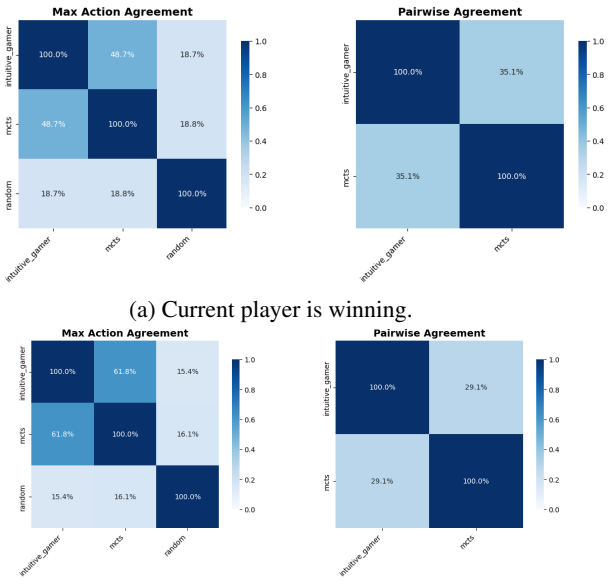
We randomly sample 10,000 game states for each game defined in Figure 1, and compare `Intuitive Gamer`, `MCTS`, and `Random`. As shown in Figure 2, `Intuitive Gamer` and `MCTS` are in agreement on the optimal action approximately 50 percent of the time, as opposed to random, which agrees with both roughly 20 percent of the time. However, when we further probe into the pairwise agreement between Intuitive Gamer and MCTS, the percent drops. This is because the nature of the value function for Intuitive Gamer causes sharp peaks and a notion of "overconfidence" around specific actions that doesn't necessarily align aside from the optimal action.

In order to gain further insight into states that cause convergence between the Intuitive Gamer value function and MCTS, we filter game states according to whether the player is winning or losing. Interestingly, as shown in Figure 3, in the states where the player is losing, the heuristic of Intuitive Gamer is able to more often identify the expert level move as compared to situations where the player is winning. This can likely be attributed to the intuition that defensive planning is easier when compared to offensive planning, which often requires a farther look ahead. However, it should be noted that in both cases the pairwise agreement is quite low between expert policy (MCTS) and intuitive gamer.

These results are important as they show that a unique notion of sensibility exists for different policies and skill levels. We subsequently build on this insight to adjust priorities within the opponent-aware framework to win more quickly when possible.

Opponent-Aware Intuitive Gamer Framework

In this section, we present an extension to the work in (Collins et al., 2025) to capture human's innate ability to reason about



(a) Current player is winning.

(b) Current player is losing.

Figure 3: Alignment between policies over only states in $\mathcal{S}$ where current player is either winning (top) / losing (bottom).

opponent's skill level and accordingly adjust their overall strategy. As discussed above the Intuitive Gamer model is a composition of three underlying heuristics that balance offense, defense, and general auxiliary values.

$$\mathcal{V}(s,a) = \mathbf{w} \cdot U(s,a)^T$$

where $\mathbf{w} = \begin{bmatrix} w_{\texttt{self}} & w_{\texttt{opp}} & w_{\texttt{aux}} \end{bmatrix}$ and $U(s,a) = \begin{bmatrix} U_{\texttt{self}}(s,a) & U_{\texttt{opp}}(s,a) & U_{\texttt{aux}}(s,a) \end{bmatrix}$

However, as we postulated earlier, humans dynamically adjust these priorities (even for new games) based on their estimate of their opponents skill level. Consequently, we propose that rather than having a static $\mathbf{w}$, we can calculate $\mathbf{w}$ conditioned on $P(\pi_{\texttt{opp}} \mid \tau)$ as follows:

$$\hat{\mathbf{w}}(\tau) = \sum_{\pi \in \Pi} P(\pi \mid \tau) \mathbf{w}^*_\pi \tag{4}$$

where $\forall \pi \in \Pi$ we empirically find an optimal $\mathbf{w}^*_\pi$ to most efficiently beat an opponent with policy π.

Optimal Weight Sweep

In order to implement our dynamic $\hat{\mathbf{w}}(\tau)$ we need to empirically determine how the weights need to be adjusted. We perform a weight sweep from $[0,2]$ with 0.2 increments. For each weight candidate, we rollout 50 games where we compute the win percentage ω and the average number of turns to win T_w. Given that we define optimal as the weight combination that wins the most often and the fastest, we define the following metric which discounts the number of turns to win according to the win rate as follows:

$$\bar{T}_w = \frac{T_w}{\omega^\beta}$$

where β controls how much we penalize the win percentage being low. We subsequently define $\mathbf{w}^*_\pi$ for each opponent policy π that minimizes $\bar{T}_w$. We have included the resulting weights in Table 1.

Opponent Policy (π)	$\mathbf{w}^*_{\texttt{self}}$	$\mathbf{w}^*_{\texttt{opp}}$	$\mathbf{w}^*_{\texttt{aux}}$
`Random Policy`	1.5	0.3	1.5
`Intuitive Gamer`	0.9	0.3	0.3
`MCTS (Expert)`	1.5	0.9	1.5

Table 1: Empirically determined optimal weights $\mathbf{w}^*_\pi$ for the Opponent-Aware Intuitive Gamer model when facing different opponent policies π. Weights are optimized to minimize the cost metric $\bar{T}_w$.

Interestingly, the results align with prior intuition. When competing against a more novice player you care much more about your own objective to win rather than defending against the other person. On the other hand, when competing with MCTS (Expert), we need to prioritize defense as well. It is important to note that with MCTS (Expert), even with the above weights, the max win rate was merely 27 percent which is expected.

Opponent Aware Intuitive Gamer Overview

Algorithm 1 describes our full approach for selecting the next action at a given timestep to exploit the opponent using some base policy π_{opp}. First, we compute the sensibility score of the full opponent trajectory τ (updated to include the latest action) using either of our two proposed methods in 2 and 3 (lines 1-4). After computing the softmax over the sensibility values, we obtain a posterior distribution $\mathcal{P}_t$ over all policies in the policy bank, $\pi \in \Pi$ (lines 5-8). An alternative way to implement this would be to update the previous posterior given a new pair (s_{t-1}, a_t). However, we effectively accomplish the same posterior update by recomputing $P(\cdot|\pi)$ with the whole opponent (state, action) list τ_t. After we compute the opponent posterior, we update the weights of our value function, $\hat{\mathrm{w}}$, by taking the weighted average of the posterior probabilities and the empirically derived optimal weights for each policy w^*_π (lines 9-12). Then, we calculate the value function weighted by $\hat{\mathrm{w}}$ at each valid next state s_{t+1}, compute the softmax to create a valid distribution over likelihoods, and finally sample an action in proportion to the likelihood distribution (lines 13-19).

Algorithm 1: Opponent-Aware Intuitive Gamer

Input : Current state s_t,
Opponent previous move $(s_{t-1}, a_{\text{opp}})$,
Policy bank Π, optimal weights set
$\mathbf{W}^* = \{\mathbf{w}^*_\pi \mid \pi \in \Pi\}$
Previous belief distribution $\mathcal{P}_{t-1}(\pi)$
Smoothing factor α
Output : Action a_t, updated belief $\mathcal{P}_t(\pi)$

1 $\tau_t \leftarrow \tau_{t-1} \cup \{(s_{t-1}, a_{\text{opp}})\}$
2 **for** $\pi \in \Pi$ **do**
3 $\quad P(\tau_t|\pi) \leftarrow \text{ComputeSensibility}(\tau_t, \pi)$
4 **end for**
5 $Z \leftarrow \sum_{\pi' \in \Pi} \exp(P(\pi'|\tau))$;
6 **for** $\pi \in \Pi$ **do**
7 $\quad \mathcal{P}_t(\pi) \leftarrow \frac{\exp(P(\pi|\tau)}{Z}$;
8 **end for**
9 $\hat{\mathbf{w}} \leftarrow \mathbf{0}$;
10 **for** $\pi \in \Pi$ **do**
11 $\quad \hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \mathcal{P}_t(\pi) \cdot \mathbf{w}^*_\pi$;
12 **end for**
13 **for** $a \in \mathit{ValidActions}(s_t)$ **do**
14 $\quad U(s_t, a) \leftarrow [U_{\texttt{self}}(s_t, a), U_{\texttt{opp}}(s_t, a), U_{\texttt{aux}}(s_t, a)]$;
15 $\quad V[a] \leftarrow \hat{\mathbf{w}} \cdot U(s_t, a)^T$;
16 **end for**
17 $Z_v \leftarrow \sum_a \exp(V[a])$;
18 Sample $a_t \sim \frac{\exp(V[a])}{Z_v}$
19 **return** $a_t, \mathcal{P}_t$

Results

In order to evaluate the effectiveness of our opponent aware intuitive gamer, we compare the dynamically updating priority framework with static prior weight definitions. Unfortunately, as shown in Figure 4, there are no statistically significant improvements in the speed at which our extension wins games. However, this might be explained by the fact that when competing against random, even the static weight frameworks win games in roughly 6 moves which gives marginal room for observable improvement. That being said, an interesting observation is that the log likelihood updating framework translates to better planning than the agreement counting. We hypothesize this is a consequence of the fact that log likelihood framework allows for stochasticity in agents' actions whereas the action agreement definition is lossy and forgoes all other information beyond the *most sensible* actions.

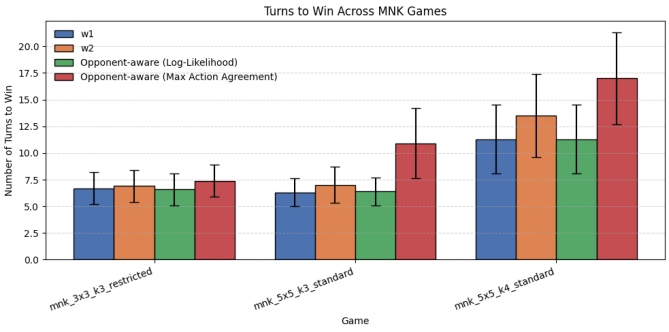
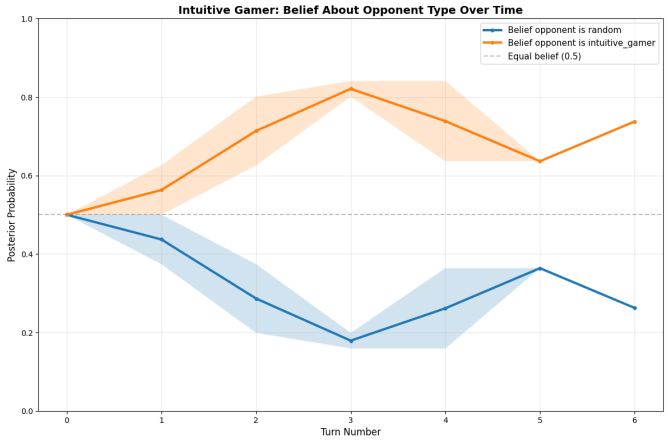
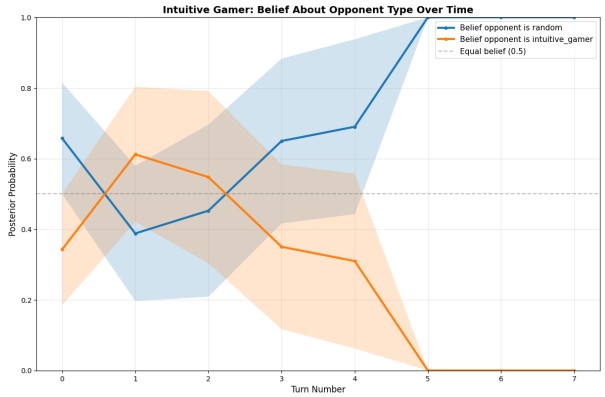


Figure 4: Average number of turns to win across game variants against Random for different priority weighting in the Intuitive Gamer Framework.

To gain further insight, we plot the posterior belief of the opponent's underlying skill level in a face-off between an intuitive gamer and a fellow intuitive gamer or novice (random) as shown in Figure 5. Despite the model being able to identify the correct opponent policy in both cases, given how short these games are, there is insufficient time to take advantage



(a) Posterior belief while competing with Intuitive Gamer.



(b) Posterior belief while competing with random (novice).

Figure 5: Maintained belief likelihoods of opponents skill levels over number of turns passed.

of this information. As seen in Figure 5b, it only converges to 100 percent likelihood on opponent being random after 5 turns, and in the case of its opponent being a fellow intuitive gamer it never fully converges.

These results suggest that there exists a better framework for updating the weights rather than a weighted average. We hypothesize that through the weighted average, the weights get stuck at a sub-optimal combination that yields poor decision making. An alternate approach would be to simply choose the optimal w for the most likely opponent skill level that explains their past moves, since weight settings might not be meaningfully composable.

We additionally conduct simulated plays across game variants between iteration-varying MCTS and depth-varying Intuitive Gamer to gain an intuition for how well a depth-varying Intuitive Gamer approximates an expert policy. The plot in Figure 6 shows that Intuitive Gamer drops in win rate when iteration count increases for MCTS, reaching equivalent performance across all depths at around 100 iterations. This is expected since MCTS conducts random play after expanding a node, so it requires some minimal iteration count to converge to reasonable estimates of future rewards. The

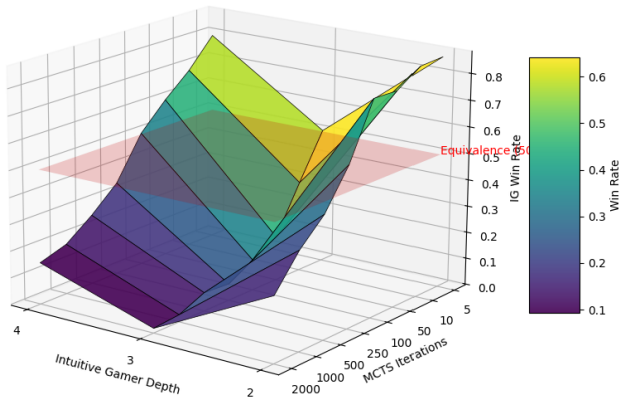


Figure 6: Win percentage across game variants between Depth Varying IG and Iteration Varying MCTS

somewhat counterintuitive result is that increasing the depth of intuitive gamer and averaging the value function at leaf nodes does not lead to a more performant policy. This is likely due to the mismatch between the average reward expectation across actions from Intuitive Gamer and the worst case action taken by an adversarial MCTS. The averaging dilutes the negative reward at seriously disadvantaged states which the expert MCTS policy will force the Intuitive Gamer into. The discrepancy at depth 3 is due to the fact that search terminates on the agent's own turn and averaging the value function at the leaf node creates an overly optimistic bias, since the agent evaluates favorable configurations without recognizing that an expert opponent would have blocked the trajectory leading to them in the previous move. Therefore, the more meaningful comparison points are at even depths.

Future Directions

This project presented an exciting opportunity to leverage Bayesian reasoning to extend a state of the art model for human cognition in novel games. Having established this foundation, several promising directions for future work emerge. It would be interesting to conduct a human study that observes humans innate ability to infer opponent's skill levels and appropriately adjust their priorities to win. Additionally, extending the update framework beyond a weight sweep would be interesting. Some alternate options include a hierarchical Bayesian framework whose underlying value function is completely different, conditioned on who we believe we are playing against. Pursuing these extensions would move us closer to computational models that not only predict human play, but mirror how humans flexibly reason about opponents in unfamiliar strategic settings.

References

Allen, K. R., Smith, K. A., Piterbarg, U., Chen, R., & Tenenbaum, J. B. (2020). Abstract strategy learning underlies flexible transfer in physical problem solving. In *Proceed-*

ings of the annual meeting of the cognitive science society (Vol. 42).

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302-29310. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1912341117 doi: 10.1073/pnas.1912341117

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.

Collins, K. M., Zhang, C. E., Wong, L., da Costa, M. B., Todd, G., Weller, A., ... Tenenbaum, J. B. (2025). People use fast, flat goal-directed simulation to reason about novel problems. *arXiv preprint arXiv:2510.11503*.

Jara-Ettinger, J., Baker, C., Ullman, T. D., & Tenenbaum, J. B. (2025). Theory of mind and inverse decision-making. In J. Tenenbaum, T. Ullman, et al. (Eds.), *Bayesian models of cognition: Reverse engineering the mind* (pp. —). Cambridge, MA, USA: MIT Press. Retrieved from https://www.tomerullman.org/papers/BBB_chapter14.pdf

Nash, J. F. (2024). Non-cooperative games. In *The foundations of price theory vol 4* (pp. 329–340). Routledge.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Silver, T., Allen, K. R., Lew, A. K., Kaelbling, L. P., & Tenenbaum, J. (2019). *Few-shot bayesian imitation learning with logical program policies.* Retrieved from https://arxiv.org/abs/1904.06317

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, *33*, 19238–19250.

Author Contribution Statement

Please note that all ideation/formulation was conducted collaboratively. The implementation details were distributed as outlined in the following statements however we both had contributions and help for the other's components.

Michal Contribution Statement

I implemented the expert Monte Carlo Tree Search (MCTS) baseline and developed the depth-limited search extension for the Intuitive Gamer model. I additionally implemented the game variants, including the logic for asymmetric win conditions and constraints. I implemented the comparative experiments between the depth-limited Intuitive Gamer and MCTS, performed the hyperparameter sweeps to derive the optimal weight configurations for the base Intuitive Gamer value function for each policy combination, and generated the visualizations characterizing move sensibility across divergent game states.

Aryan Contribution Statement

I implemented the baseline Intuitive Gamer module and developed the posterior updating and dynamic weight adjustment frameworks. I also designed and implemented the evaluation metric used for hyperparameter sweeps to identify optimal weight configurations. In addition, I built the underlying code infrastructure and configuration system supporting these components, and developed a game-agnostic sampling framework for agreement analysis that filters game states to provide deeper insight into when policies converge and diverge. Finally, I conducted comprehensive performance analyses across multiple games, including policy agreement analyses, and implemented the plotting and visualization code used for result analysis.