Improving Human Performance with Value-Aware Interventions: A Case Study in Chess

Saumik Narayanan, Raja Panjwani, Siddhartha Sen, Chien-Ju Ho

Keywords: Human-AI Interaction, Chess, Sequential Decision Making, AI Interventions

Summary

Artificial intelligence systems are increasingly used to assist humans in sequential decision-making tasks. A common approach is to recommend actions that are optimal according to a strong model of the environment. Such recommendations, however, implicitly assume that the decision maker will continue to act optimally afterward. Human decision makers often deviate from optimal play, meaning that actions that are optimal in isolation may lead to states that are difficult for humans to navigate successfully. This raises a fundamental question: how should AI systems intervene when assisting suboptimal decision makers in sequential environments? In this work, we study value-aware interventions that account for how humans are likely to behave after receiving assistance. Our approach builds on a basic principle from reinforcement learning: when a policy is suboptimal, discrepancies arise between the actions taken by the policy and those that maximize the expected value of future outcomes. We show how such discrepancies can be used to identify opportunities where interventions improve downstream performance. To operationalize this idea, we develop a data-driven framework that learns models of human decision-making from large datasets and uses these models to guide intervention decisions. We study these ideas in the domain of chess, which provides large-scale human gameplay data and strong AI systems for evaluating intervention strategies.

Contribution(s)

1. We introduce a framework for identifying value-aware intervention opportunities in sequential decision-making tasks using discrepancies between a human policy and its associated value function. In the single-intervention setting, the intervention that maximizes the human value function emerges naturally from the formulation, and we propose a tractable heuristic for settings with multiple interventions under a budget constraint.
   Context: Yu & Ho (2022) studied environment design with stylized human behavior models. Bastani et al. (2026) provided recommendations using optimal action strategies, analogous to one of our baselines.
2. This paper demonstrates that value-aware intervention strategies can be implemented using data-driven models of human decision-making and evaluates the approach through large-scale simulations and a human-subject study in the domain of chess.
   Context: Prior work evaluates interventions primarily in simplified environments. We study a complex sequential decision domain with large-scale human behavioral data, strong AI baselines, and precise notions of human skill levels, enabling both realistic simulation and human-subject evaluation.

Improving Human Performance with Value-Aware Interventions: A Case Study in Chess

Saumik Narayanan[1], **Raja Panjwani**[2,3], **Siddhartha Sen**[2], **Chien-Ju Ho**[1]
saumik@wustl.edu, rp1223@georgetown.edu, sidsen@microsoft.com, chienju.ho@wustl.edu

[1]**Washington University in St. Louis**
[2]**Microsoft Research**
[3]**Georgetown University**

Abstract

AI systems are increasingly used to assist humans in sequential decision-making tasks, yet determining when and how an AI assistant should intervene remains a fundamental challenge. A potential baseline is to recommend the optimal action according to a strong model. However, such actions assume optimal follow-up actions, which human decision makers may fail to execute, potentially reducing overall performance. In this work, we propose and study value-aware interventions, motivated by a basic principle in reinforcement learning: under the Bellman equation, the optimal policy selects actions that maximize the immediate reward plus the value function. When a decision maker follows a suboptimal policy, this policy–value consistency no longer holds, creating discrepancies between the actions taken by the policy and those that maximize the immediate reward plus the value of the next state. We show that these policy-value inconsistencies naturally identify opportunities for intervention. We formalize this problem in a Markov decision process where an AI assistant may override human actions under an intervention budget. In the single-intervention regime, we show that the optimal strategy is to recommend the action that maximizes the human value function. For settings with multiple interventions, we propose a tractable approximation that prioritizes interventions based on the magnitude of the policy–value discrepancy. We evaluate these ideas in the domain of chess by learning models of human policies and value functions from large-scale gameplay data. In simulation, our approach consistently outperforms interventions based on the strongest chess engine (Stockfish) in a wide range of settings. A within-subject human study with 20 players and 600 games further shows that our interventions significantly improve performance for low- and mid-skill players while matching expert-engine interventions for high-skill players.

1 Introduction

AI systems are increasingly used to assist humans in complex decision-making tasks, ranging from strategic planning and navigation to domains such as medicine, programming, and finance (Esteva et al., 2017; Vaithilingam et al., 2022; Fischer & Krauss, 2018). In many of these settings, decisions unfold sequentially: early choices influence the states encountered later and ultimately determine the final outcome. Designing AI systems that effectively assist humans in such environments raises a fundamental question: when and how should an AI assistant intervene in the decision process?

A natural baseline is to recommend the action that would be optimal according to a strong expert model. For example, in chess, an AI assistant could simply suggest the move preferred by a top

engine such as Stockfish. However, optimal actions are typically evaluated under the assumption that the decision maker will continue to act optimally in the future. Human decision makers, however, often deviate from optimal play. As a result, actions that are optimal under perfect continuation may lead to states that are difficult for humans to navigate, increasing the likelihood of future mistakes. Conversely, actions that appear objectively suboptimal may lead to trajectories that are easier for humans to execute successfully, ultimately yielding better overall outcomes.

This intuition can be formalized through a basic principle in reinforcement learning. Under the Bellman equation, the optimal policy selects the action that maximizes the immediate reward plus the value of the next state. In other words, the policy and the value function satisfy a consistency condition: the action chosen by the policy maximizes the expected value of future outcomes (Puterman, 1994). When a decision maker follows a suboptimal policy, this policy-value consistency no longer holds. Rather than viewing this discrepancy purely as a theoretical artifact, we use it as a signal for identifying opportunities where interventions can improve downstream performance.

Building on this idea, we study value-aware interventions for assisting suboptimal decision makers in sequential environments. Instead of recommending the action that maximizes the value under optimal play, the assistant recommends the action that maximizes the expected outcome when the trajectory continues under the human's policy. We formalize this problem as a Markov decision process in which an AI assistant may override human actions under a limited intervention budget. This formulation captures a common practical constraint: interventions may be costly due to cognitive load, reduced autonomy, or system design considerations, making selective intervention important.

A key challenge in applying this approach is that the human policy and its associated value function are generally unknown. In this work, we address this challenge using a data-driven approach. Leveraging large datasets of human decision trajectories, we learn models that approximate both how humans act and how their actions translate into expected outcomes. These learned models allow us to estimate policy-value discrepancies and identify states where interventions are most likely to improve performance.

We study this framework in the domain of chess, which provides a particularly useful testbed for human-AI decision support, with several key advantages: First, large-scale datasets of human gameplay are available, enabling accurate modeling of human policies across a wide range of skill levels (Lichess, 2026). Second, chess has extremely strong AI systems that can approximate optimal play, providing reliable baselines for comparison (Stockfish developers, 2026). Third, recent work has demonstrated that neural networks can learn models that closely mimic human decision-making patterns, enabling realistic simulations of how humans are likely to act in future positions (McIlroy-Young et al., 2020). Together, these properties make chess an ideal environment for studying how AI assistance should adapt to human behavior in sequential tasks.

Using large-scale datasets of online chess games, we train behavioral cloning models that estimate both human policies and their associated value functions across different skill levels. With these models, we analyze intervention strategies under both single- and multiple-intervention settings. Our results show that interventions designed to maximize the expected outcome under human play can substantially improve performance when interventions are limited. In particular, we find that these interventions outperform recommendations based purely on the strongest chess engine across all skill levels when the number of interventions is small. We further validate these findings through a human-subject experiment involving 20 players and 600 games, showing that interventions significantly improve performance for low-skill and medium-skill players while matching expert interventions for high-skill players.

Overall, this work makes the following contributions. First, we introduce a framework for identifying intervention opportunities in sequential decision-making tasks using discrepancies between human policies and value functions. Second, we derive value-aware intervention strategies under both single-intervention and multiple-intervention settings. Third, we demonstrate that such strategies can be implemented in practice using data-driven models of human behavior and validate their effectiveness through large-scale simulations and a human-subject study in the domain of chess.

2 Related Work

Human Interventions in Sequential Environments. Our work most closely relates to a line of work on sequential interventions on human decision making, which looks at how to best intervene on a human's trajectory of decisions in order to improve their performance. Yu & Ho (2022) formulate this as an environment design problem and propose to either update the environment rewards or make "action nudges" to change human behavior under an intervention budget, and show that human-aware interventions improve performance in a gridworld environment under theoretical models of human behavior. Our approach directly builds on this work, by showing how to apply this framework in real-world domains using data-driven models of human behavior.

A different approach is taken by Bastani et al. (2026), who show that, given human trajectories in an Overcooked environment, selecting the intervention with the highest expected difference between the human and an optimal agent leads to improved performance. However, their approach does not account for the impact of suboptimal human policies on the estimation of value and Q functions. One of the main contributions of our approach, in contrast, is the use of human value-aware interventions that explicitly account for how humans are expected to perform after the intervention, rather than assuming optimal human behavior. Moreover, their method relies on the (limited) observed trajectories to empirically estimate Q functions, whereas we leverage data-driven methods to estimate both the policy and value functions.

Alternative notions of interventions can be seen in Nofshin et al. (2024), who modify the behavioral clone's MDP parameters (e.g. discount factor, costs, rewards), rather than their actions, and in Straitouri et al. (2025), who reduce the action space rather than directly selecting actions. Another line of work looks at the effect of AI recommendations (rather than direct interventions) on human decision making, including in sequential decision making settings. For instance, Chen et al. (2023) and Grand-Clément & Pauphilet (2026) both analyze settings where AI recommendations are not fully followed, and learn to provide recommendations which are robust to partial adherence. More broadly, our work also connects to behavior-aware algorithm design, where models of human decision-making are incorporated into the optimization problem rather than treated as exogenous; for example, Yu et al. (2023) study how to encode human behavior models in information design.

Modeling Human Behavior. There are a variety of approaches to modeling human behavior in sequential decision making settings. We use behavioral cloning (Bain & Sammut, 1995), a method to learn a supervised model from data which is widely used in domains such as robotics (Florence et al., 2022) and autonomous driving (Wang et al., 2022). Inverse Reinforcement Learning is an alternative approach, which learns the underlying reward function of human behavior, rather than learning the policy directly (Ng et al., 2000).

Specifically in chess, there have been a number of recent works which have looked at modeling human behavior. McIlroy-Young et al. (2020) show that a behavior cloning approach over a large human dataset can successfully predict actions, and this has been followed up with works showing that individualized models can even better predict behavior when accounting for playing style (McIlroy-Young et al., 2022). Other works have looked at adding search to further improve modeling performance (Jacob et al., 2022; Zhang et al., 2024), and Narayanan et al. (2022) studied approaches to learning strong models with very small human data budgets.

3 Setting and Methodology

3.1 Problem Setting

Our work is situated in the context of a Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mathcal{S}_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}(s)$ is the action space, $\mathcal{P}(s, a)$ is the transition function, $\mathcal{R}(s, a)$ is the reward function, γ is the discount factor, and $\mathcal{S}_0$ is the distribution of starting states. In this paper, we focus on finite MDPs with $\gamma = 1$ and episode length T, though our discussion generalizes to other MDPs as well.

A (stochastic) policy is denoted by $\pi(a|s)$. Let $\zeta = \{(s_t, a_t, r_t)\}_{t=1}^{T}$ denote a trajectory generated by policy π starting from an initial state s_0. The value function of a policy is defined as $V^{\pi}(s) = \mathbb{E}_{\zeta \sim \mathcal{Z}^{\pi}(s)}\left[\sum_{t=1}^{T} r_t\right]$, where $\mathcal{Z}^{\pi}(s)$ denotes the distribution over trajectories induced by policy π starting from state s. The expected reward of a policy is $J(\pi) = \mathbb{E}_{s_0 \sim \mathcal{S}_0}\left[V^{\pi}(s_0)\right]$. The Q-function of a policy is defined as $Q^{\pi}(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}\left[\mathcal{R}(s,a) + V^{\pi}(s')\right]$.

In this work, our goal is to design an intervention policy that improves the expected performance of the human decision maker while limiting the frequency of interventions. We assume access to a human policy π_H, and our task is to learn an *intervention policy* consisting of (i) a *gating function* $\phi(s) \in [0,1]$ that denotes the probability of overriding the human at state s, and (ii) an *override policy* $\pi_I(a|s)$ that specifies the action taken when an intervention occurs. We use $\pi_H \oplus (\phi, \pi_I)$ to denote the policy after intervening human policy π_H with intervention (ϕ, π_I):

$$(\pi_H \oplus (\phi, \pi_I))(a|s) = \begin{cases} \pi_I(a|s) & \text{with probability } \phi(s) \\ \pi_H(a|s) & \text{with probability } 1 - \phi(s). \end{cases} \tag{1}$$

Under the constraint that the expected intervention rate lies under some budget frequency $B \in [0,1]$, denoting the maximum expected fraction of timesteps at which the AI may override the human policy, our goal is to find the optimal intervention policy and gating function:

$$\arg\max_{\phi, \pi_I} J(\pi_H \oplus (\phi, \pi_I)) \quad \text{s.t.} \quad \frac{1}{T}\,\mathbb{E}_{\zeta \sim \mathcal{Z}^{\pi_H \oplus (\phi, \pi_I)}}\left[\sum_{t=1}^{T} \phi(s_t)\right] \leq B. \tag{2}$$

3.1.1 A Case Study in Chess

In this work we focus on the domain of chess as a case study. Chess provides a rich sequential environment with well-defined rules, large-scale datasets of human play, and strong AI systems that allow us to analyze intervention strategies in a realistic setting. In chess, the elements of the MDP correspond naturally to the game structure. A state $s \in \mathcal{S}$ corresponds to a board position, an action $a \in \mathcal{A}(s)$ corresponds to a legal move available in that position, and the transition function $\mathcal{P}(s,a)$ is determined by the rules of chess. The reward function reflects the final game outcome, typically defined as 1 for a win, 0.5 for a draw, and 0 for a loss from the perspective of the player.

A key component of our framework is the human policy $\pi_H(a|s)$ and its associated value function $V^{\pi_H}(s)$, which capture how a human player behaves and the expected outcome when play continues according to that behavior. In practice, we do not have direct access to these quantities. Instead, we learn an approximation using behavioral cloning trained on large datasets of human games, using the methodology by McIlroy-Young et al. (2020). The behavioral cloning model provides (i) a policy head that predicts the distribution of human moves $\pi_H(a|s)$ and (ii) a value head that estimates the expected outcome when the trajectory continues under the learned human policy, providing an estimate of $V^{\pi_H}(s)$. For comparison with optimal play, we use the chess engine Stockfish to approximate the optimal policy π^* and the associated optimal value function $V^{\pi^*}(s)$.

3.2 Value-Aware Intervention Design

We now describe our approach for designing value-aware interventions. The key idea is that when assisting a human decision maker in a sequential decision-making environment, the assistant should not necessarily recommend the action from the optimal policy. Actions that are optimal under perfect play may lead to states that are difficult for humans to navigate, resulting in mistakes later in the trajectory. Instead, the assistant should recommend the action that maximizes the expected downstream performance of the human who will continue the trajectory. As described in Section 3.1.1, we approximate the human policy $\pi_H(a|s)$ and the associated value function $V^{\pi_H}(s)$ using a behavioral cloning model trained on human gameplay data.

In reinforcement learning, optimal policies satisfy the Bellman optimality condition, which states that the optimal action maximizes the expected immediate reward plus the value of the next state. As

a result, the optimal policy $\pi^*(a|s)$ selects actions that maximize $Q^{\pi^*}(s,a)$, ensuring consistency between the policy and the value function. Human decision makers, however, often follow suboptimal policies, and therefore there can be discrepancies between the actions by the human policy and those that would maximize the expected value of future outcomes. We use these discrepancies to identify opportunities for intervention. Rather than selecting actions that lead to the objectively highest-value states (as done by strong chess engines such as Stockfish), interventions should be chosen based on how they affect the expected outcome when the human continues the trajectory. Formally, this corresponds to evaluating candidate actions using the human value function V^{π_H} and the associated action-value function Q^{π_H}.

3.2.1 Single-Intervention Regime

First, we consider the setting where at most one intervention is allowed per trajectory. In this setting, there is a simple solution for the optimal intervention when we have access to π_H, and V^{π_H}. [1]

When at most one intervention is allowed per episode, the post-intervention behavior follows π_H by definition. Consequently, the value of taking action a at state s is given by $Q^{\pi_H}(s,a)$, since the remainder of the trajectory follows the human policy. In this regime, the benefit of overriding the human at state s with action a is

$$\Delta^{\pi_H}(s,a) = Q^{\pi_H}(s,a) - V^{\pi_H}(s),$$

which measures the improvement in expected outcome relative to allowing the human to act according to their own policy. The optimal override action at state s for policy π_H is therefore $a(\pi_H, s) = \arg\max_a Q^{\pi_H}(s,a)$. Similarly, over the course of a trajectory ζ generated by π_H, the optimal state to intervene is the state that maximizes this improvement: $s(\pi_H, \zeta) = \arg\max_s \left[\max_a \Delta^{\pi_H}(s,a)\right]$.

Thus, in the single-intervention setting, Q^{π_H} determines both (i) the best action to recommend and (ii) the optimal criterion for selecting the intervention state. Given a full human trajectory ζ, the optimal intervention (ϕ^*, π_I^*) that solves Equation 2 satisfies that (1) $\phi^*(s) = 1$ if $s = s(\pi_H, \zeta)$ and $\phi^*(s) = 0$ otherwise, and (2) $\pi_I^*(a|s) = 1$ if $a = a(\pi_H, s)$ and $\pi_I^*(a|s) = 0$ otherwise.

In practice, however, the realized trajectory is typically not known in advance, and the assistant must make intervention decisions online as states are encountered. This introduces an additional timing problem: intervening prematurely may use the limited intervention opportunity on a state with modest benefit, whereas delaying intervention may forgo the opportunity to act at a high-value state. Thus, in the online setting, the assistant must compare the immediate expected gain from intervening at the current state with the continuation value of preserving the intervention opportunity for future states. This formulation corresponds to a standard optimal stopping problem (Peskir & Shiryaev, 2006).

3.2.2 Multiple-Intervention Regime

The above intervention methodology is optimal when at most one intervention is allowed per trajectory. When multiple interventions are permitted, however, selecting actions based on $Q^{\pi_H}(s,a)$ is no longer guaranteed to be optimal. The reason is that Q^{π_H} assumes the remainder of the trajectory follows the human policy π_H, whereas in a multiple-intervention setting additional interventions may occur later, causing the resulting policy to deviate from π_H.

Ideally, the value of an action should therefore be evaluated under the post-intervention policy $\bar{\pi} = \pi_H \oplus (\phi, \pi_I)$. This leads to a significantly more challenging optimization problem than in the single-intervention setting, since the value function $Q^{\bar{\pi}}$ depends on the intervention policy (ϕ, π_I) itself. To obtain a tractable procedure, we approximate $Q^{\bar{\pi}}$ using Q^{π_H}. This approximation is reasonable when the intervention budget B is relatively small. When interventions occur infrequently, the resulting trajectory distribution remains close to the human trajectory distribution induced by

[1] If we have access to value function V^{π_H}, we can derive Q^{π_H} by adding the immediate reward.

π_H. Under this approximation, we evaluate intervention opportunities using the same improvement signal as in the single-intervention setting, $\Delta^{\pi_H}(s,a) = Q^{\pi_H}(s,a) - V^{\pi_H}(s)$. This yields the simplified optimization problem. The override policy π_I^* follows the same solution as in the single-intervention regime: $\pi_I^*(a|s) = 1$ if $a = a(\pi_H, s)$ and $\pi_I^*(a|s) = 0$ otherwise. The gating function ϕ^* is obtained by solving

$$\arg\max_{\phi} \; \mathbb{E}_{s\sim\zeta^{\pi_H}}\left[\phi(s)\max_a \Delta^{\pi_H}(s,a)\right] \quad \text{s.t. } \mathbb{E}_{s\sim\zeta^{\pi_H}}[\phi(s)] \leq B.$$

Although this approach relies on approximating $Q^{\bar{\pi}}$ with Q^{π_H}, which is most accurate when B is small, we find empirically that it performs well even for larger intervention budgets. In our experiments, the method continues to outperform baseline policies even when we allow interventions up to 50% of the time.

4 Simulation Results

In the previous section, we introduced a framework for value-aware interventions and derived intervention strategies for both the single-intervention and multiple-intervention regimes. In particular, our analysis shows that when only a single intervention is allowed, selecting the action that maximizes the human value function Q^{π_H} naturally emerges as the optimal intervention. For settings with multiple interventions, we proposed a practical approximation that prioritizes interventions based on the discrepancy between the human policy and its associated value function.

In this section, we empirically evaluate these predictions through a large-scale simulation study in the domain of chess. Chess provides a rich sequential environment with strong AI baselines and large datasets of human gameplay, allowing us to model human behavior and evaluate intervention strategies. In particular, we compare interventions that optimize the expected outcome under the human policy with interventions that select actions based on a strong expert model (Stockfish), which assumes optimal continuation. Using models of human behavior trained on large datasets of online games, we simulate both human play and AI interventions to measure how different strategies affect downstream performance. We evaluate these strategies in both the single-intervention setting and multiple-intervention regimes under limited intervention budgets.

4.1 Experiment Setup

For all experiments in this paper, we use a Behavioral Cloning (BC) approach to model human behavior. First, we sample a dataset of 256 million chess positions, actions, and game outcomes from the Lichess dataset (Lichess, 2026), such that the blitz rating of the players range uniformly between 400 and 2800. We then fine-tune an existing pretrained Leela T82 model on this dataset. Notably, we train a single BC model across all ratings, but include the players' ratings as an input to the model, allowing it to learn a single model which can generalize across all ratings. As a result, we are able to use this BC model to predict both the human policy function π_H and the human value function V^{π_H} at any given rating. [2]

4.2 Single-Intervention Experiments

To compute the efficacy of single interventions, we leverage the Lichess dataset of human games and simulate interventions at random positions sampled from the dataset. We then evaluated the counterfactual outcome after intervening at these positions, and then, we simulate 64 game rollouts from the resulting position using the trained BC model as our approximation of the human policy, setting the policy rating for both sides as the player rating at the initial position.

Then, we evaluated three different classes of intervention strategies in each position:

[2] We provide more details on model training in Appendix A, and the full code to reproduce our results is available at https://github.com/saumikn/chess_training.

Human Move: This is the actual move that the human played in the game, and serves as a baseline for comparison.

Stockfish Intervention: We intervene with the move which maximizes the expected value according to Stockfish (e.g. the optimal move assuming optimal follow-up).

Valuemax Intervention: We intervene with the move which maximizes the expected value prediction according to our trained BC model at the human's rating (e.g. the optimal move, assuming human follow-up).

We evaluate each of these strategies at 100k random positions sampled from games played by players at 800, 1200, 1600, 2000, and 2400 ratings (500k positions in total), and report the resulting win rate and standard error of each intervention strategy in Figure 1.

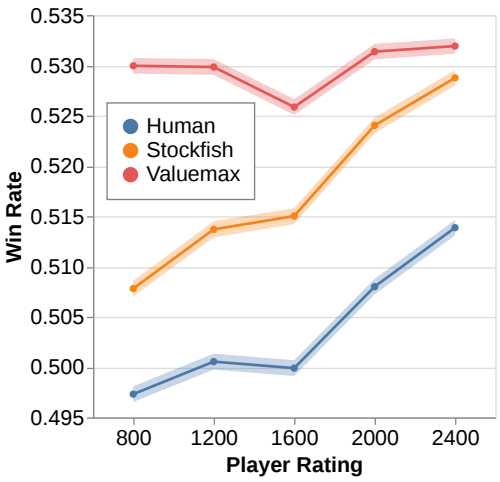

Figure 1: Win rate and standard error of each single-intervention strategy across ratings.

Overall, these results provide empirical validation for our theoretical predictions from Section 3: conditional on intervening at a given state, Valuemax consistently outperforms both the human move and the Stockfish intervention. At all ratings, the Valuemax strategy outperforms both the baseline human move and the Stockfish strategy ($p < 0.001$). Interestingly enough, the gap between Stockfish and Valuemax tends to strongly decrease as player strength goes up – for 800-rated players, Valuemax performs over 2% better than Stockfish, while for 2400-rated players, Valuemax only helps by around 0.3%. This trend aligns with intuitions, since Valuemax is explicitly exploiting the non-optimality of the human in order to make recommendations. As the human policy tends closer towards perfectly optimal, there is a smaller difference to exploit, so the difference between Valuemax and Stockfish will correspondingly reduce.

4.3 Multiple-Intervention Experiments

Next, we consider an intervention regime where we can intervene multiple times during the course of a game. This setting is more practical for real-world deployment, as we may want to provide ongoing assistance to players throughout a game, rather than just on a single move. While our theoretical framework does not provide us with optimality guarantees in this regime, we can still test the extent to which our framework described in Section 3.2.2 provides empirical improvements in performance over baselines.

In particular, we consider a threshold-based intervention strategy, where we intervene whenever the expected improvement in performance exceeds a certain threshold. This allows us to control the frequency of interventions, using higher thresholds to ensure fewer interventions and lower thresholds for more frequent interventions. At the extreme ends of the spectrum, a threshold of 1 corresponds to a 0% intervention frequency, while a threshold of 0 corresponds to a 100% intervention frequency.

Then, we evaluate two intervention strategies, Valuemax and Stockfish across a range of thresholds. Intuitively, we would expect Valuemax to outperform Stockfish when interventions are infrequent, as Valuemax is designed to maximize performance of a policy which acts similarly to an actual human at that skill level. However, as the rate of interventions increases, the more the future policy diverges from the human policy, and the less well we would expect the Valuemax strategy to be calibrated to the resulting mixed policy.

To evaluate each intervention strategy, we simulate full games starting from the initial chess position. At each move, we compute the improvement over the expected human policy (estimated using our learned BC model π_H). We then intervene if the expected improvement exceeds the specified threshold and otherwise play the move sampled from the BC model, continuing this process until the game ends. After the game, we record the overall outcome and the total number of interventions made, as a percentage of intervention opportunities. We play 2560 randomized games for each strategy, threshold, and player rating combination, and report the average intervention frequency and win rate in Figure 2.

Overall, our results align with our expectations on the relationship between intervention frequency and performance for both strategies. At each player rating, Valuemax outperforms Stockfish at small intervention budgets. For instance, for 800-rated players, Valuemax outperforms Stockfish with budgets as high as $B = 0.5$, while for 2400-rated players, this is true for budgets up to $B = 0.05$. Conversely, Stockfish outperforms Valuemax for players at each rating with large intervention budgets, as the divergence from the human policy reduces the calibration of Valuemax.

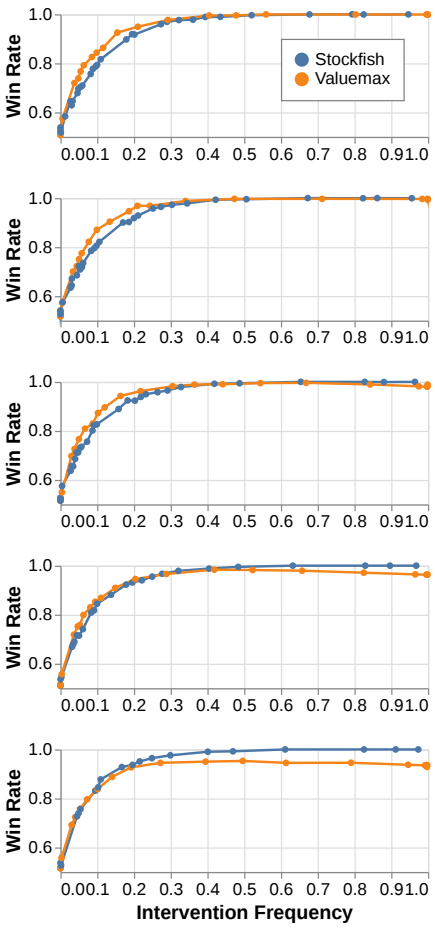

Figure 2: Win rate of each multiple-intervention strategy across player ratings and intervention frequency budgets.

4.4 Exploring Human-Understandable Concepts

One advantage of studying interventions in the domain of chess is that positions can be evaluated using a large number of structured, human-understandable concepts derived from classical chess evaluation functions. Modern engines such as Stockfish compute dozens of intermediate features that capture properties of a position such as material balance, piece activity, pawn structure, and king safety. While our intervention strategy itself does not directly use these concepts, they provide a useful lens for examining the types of positions in which interventions occur.

We conduct a limited analysis exploring how these concepts correlate with our proposed intervention strategy, and present the results in Appendix B. Here, we highlight one preliminary finding – a weak enemy king correlates very strongly with low-skill interventions, but not with high-skill interventions. One possible explanation of this is that low-skill players need help to checkmate the king while on the attack, while high-skill players already know how to deliver mate, so interventions may not be needed in those positions.

This type of analysis combining value-aware interventions and human-understandable concepts has strong potential to lead to more explainable AI assistants or even insights on human pedagogy (finding positions and concepts where a particular human is weakest). We leave this avenue of analysis to future researchers.

5 Human Study

In Section 3, we provided theoretical justifications for the Valuemax intervention strategy, and in Section 4, we validated through simulation our approach in both single-intervention and multi-

intervention regimes. However, our simulated results still rely on evaluations using a synthetic model, which leaves open the question of whether these results hold for real human players.

In this section, we aim to demonstrate using a within-subjects human experiment[3] that interventions using the Valuemax strategy indeed outperform both human policy baselines and interventions which assume that humans will follow-up with optimal behavior.

5.1 Human Participant Recruitment

By advertising at local chess clubs, we recruited 20 players to take part in our experiment. In order to ensure that participants were chess players, and to give them opponents at roughly the same skill level, we required each participant to have played at least 50 lifetime games on the Lichess or Chess.com online platforms in the Bullet, Blitz, or Rapid time controls. Because each website and each time control has a different rating scale and distribution, rating standardization was required to fairly compare players. To do this, we used the cross-platform rating dataset collected by Jensen (2025). In Table 1, we summarize the standardized ratings of the players in the study. Each participant spent approximately 60 minutes to complete our study, and we compensated each player with a $30 Amazon gift card upon completion.

Standardized Rating	**n**
800–1600	7
1600–2000	6
2000–2400	7
Total	**20**

Table 1: Summary rating statistics of human participants in our study.

5.2 Study Design

In this study, we evaluate three different single-intervention approaches: Human, Stockfish, and Valuemax. While it would be ideal to evaluate the performance of each intervention strategy across the true distribution of states encountered by players at each rating, this would require an infeasible number of games to be played by human participants, due to the small expected differences in performance between the intervention strategies. For example, in our simulated single-intervention results in Section 4, we see that the expected improvement of Valuemax over Stockfish interventions hovers between 0.3% and 2.5% across ratings, meaning that we could need tens of thousands of games to have enough statistical power to detect significant differences between the strategies.

Instead, we take a more targeted approach to evaluate the performance of each intervention strategy in states where we would expect to see significant differences in performance between the strategies. In particular, we use our simulated results to identify states where Valuemax interventions are expected to yield significantly higher improvements over Stockfish moves and human moves, and then evaluate the real-world performance of each intervention strategy in these states with human participants. This allows us to test whether the expected improvements predicted by our simulated results translate to real-world improvements when intervening with human players. In particular, we randomly selected 20 positions at each rating where the expected improvement of Valuemax over Stockfish and the BC policy[4] were each at least 20%, giving us 100 total positions.

Each participant in the user study was randomly assigned 30 positions to evaluate based on their rating, taken from position sets at the two closest ratings to them. For example, if a user had a standardized rating of 1900, we combined the 20 positions at the 1600 level and the 20 positions at the 2000 level, into a pool of 40 positions and then randomly sampled 30 positions to provide to the user.[5] Then, for each position, we randomly assigned one of the three treatments (Human, Stockfish, Valuemax), such that each participant ended up playing 10 games under each treatment

[3] Our experiment design was approved by our institution's IRB.

[4] Comparing Valuemax to the policy is a more fair comparison here than comparing to the Human move, since in many setups, we may not have access to the human move before we must decide to give a recommendation.

[5] This additional buffer allowed us to fill the full quota even if a few games were cut short due to user disconnections.

In this work, we are primarily focused on understanding which strategy leads to the best outcome, assuming that recommendations were taken consistently. Therefore, in our study design, we presented players with the board position *following* the recommendation move, having the user continue the rest of the game under a standard blitz time control of 3+2, playing against the BC policy model at the position rating.

5.3 Results

Overall, we collected data on 600 total games played by 20 human participants across the three different intervention strategies, using a custom chess platform designed for this experiment. We report the win rate for each intervention strategy across player ratings in Figure 3.

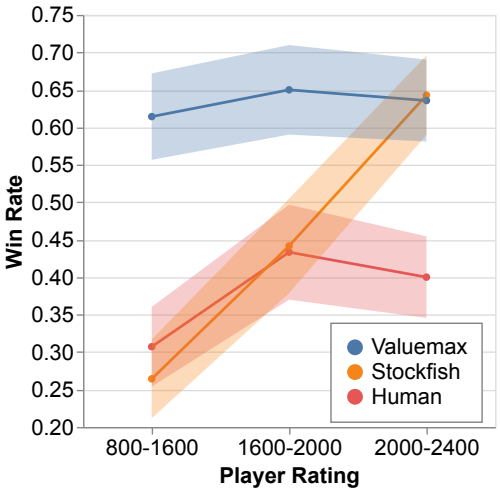

Figure 3: Human performance of each intervention strategy across ratings.

We find strong evidence that a Valuemax intervention strategy can outperform Stockfish-based interventions and the human baseline, especially at lower ratings. While we cannot directly compare the results to those in Figure 1 due to the difference in initial state distributions, the results still align with our expectations. Using a Mixed-Effects Linear Model statistical test, Valuemax significantly outperforms both Stockfish and the human baseline at all ratings under 2000 ($p < 0.001$). Interestingly, we see no effect between Valuemax and Stockfish at the highest rating levels, despite our initial expectations estimating over a 20% performance gap between Valuemax and Stockfish at every rating level. On the other hand, at the lowest skill level, we see differences greater than 35%, despite still predicting the same 20% expected performance gap. Overall, the sharp slope upwards on the Stockfish line indicates that our simulated results might underestimate the Stockfish intervention strategy at high levels, but possibly overestimate the Stockfish intervention strategy at low skill levels.

6 Conclusion and Discussion

In this work, we study how AI systems should intervene when assisting humans in sequential decision-making tasks. While a natural baseline is to recommend the action from a strong expert model, such recommendations implicitly assume optimal continuation. When human decision makers deviate from optimal play, these recommendations may lead to states that are difficult for them to navigate, reducing overall performance. We propose value-aware interventions that instead select actions based on the expected outcome when the trajectory continues under the human policy. Our analysis shows that when at most one intervention is allowed, recommending the action that maximizes the human value function naturally arises as the optimal intervention. For settings with multiple interventions, we propose a practical approximation that prioritizes interventions based on discrepancies between the human policy and value functions.

We evaluate these ideas in the domain of chess using data-driven models of human behavior trained on large datasets of gameplay. In both simulation and a human study, interventions that account for human behavior outperform recommendations based solely on optimal play when interventions are limited, particularly for low- and mid-skill players. These results support the intuition that actions optimal under perfect play may not be optimal when the decision maker follows a different policy. We also observe that the advantage of human-aware interventions depends on the intervention budget: when interventions are rare, accounting for human behavior yields clear improvements, whereas

as the intervention frequency increases the resulting trajectories diverge from the human policy, making optimal-play recommendations increasingly competitive. Finally, we conduct a preliminary analysis exploring correlations between value-aware interventions and human-understandable concepts, with potential connections to explainable AI assistance and human pedagogy.

Limitations and Future Work Our work has several limitations. First, the approach relies on learned behavioral models of human decision making, which may not perfectly capture real behavior, especially in domains with more limited data. For instance, other work (McIlroy-Young et al., 2020) found that learning models of high-skill human chess behavior can be especially difficult, and as a result, the Valuemax method may underperform at the highest skill levels compared to other baselines such as Stockfish. In general, we would expect that the better the human model is, the better our approach will perform, and additional optimizations beyond skill such as personalization would improve performance even more.

Second, our simulation and human-subject experiments rely on several approximations. Our simulation experiments depend on rollouts from the behavioral cloning model rather than real human trajectories. Our human-subject study assumes that recommendations are followed exactly, whereas in practice users may ignore or partially follow AI suggestions. The human-subject study also involved a limited number of participants and deliberately selected positions where the model predicted a large advantage for Valuemax. Thus, while the study tests whether the predicted mechanism transfers to real players, a larger-scale experiment with more naturally sampled positions would provide a stronger estimate of deployment-level performance.

Third, our experiments do not fully address the problem of *when* to take interventions. In the single-intervention experiments, we evaluate action selection conditional on intervening at a sampled position, rather than an online policy for deciding when to intervene. In the multiple-intervention experiments, the threshold-based policy controls intervention frequency in expectation but does not solve the full constrained optimization problem in Equation 2. Thus, our multiple-intervention method should be viewed as a heuristic, and jointly optimizing timing and action selection remains an important direction for future work.

Finally, our evaluation focuses on chess, a domain with abundant data, well-defined outcomes, and strong AI baselines. Future work could explore extending these ideas to other sequential decision-making settings, especially domains where human behavior is more heterogeneous, data are more limited, or interventions may affect learning and skill development over time. It would also be valuable to study interventions in continuous and real-time domains, where the assistant must decide not only what action to recommend, but also when and how frequently to provide assistance.

An exciting direction for future work is connecting our study to human pedagogy. It is plausible that intervention strategies that improve immediate performance are related to teaching strategies that improve human skill over time. However, the greedy nature of our intervention strategy may hinder long-term skill development by preventing learners from encountering critical learning opportunities. Future work could study how to design interventions that balance immediate assistance with allowing users opportunities for exploration, practice, and learning. More broadly, this highlights a tradeoff between assistance and autonomy. Future work should study how intervention frequency, timing, and explanation affect not only immediate performance, but also users' future behavior and learning.

Overall, our findings suggest that AI assistants in sequential environments should account for how users are likely to act after receiving assistance, rather than relying solely on optimal-action recommendations. This perspective is especially important when users are systematically suboptimal or when expert-recommended actions lead to states that are difficult for them to handle. Incorporating models of human behavior into the design of intervention strategies may therefore play an important role in building AI systems that more effectively support human decision making.

References

Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.

Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human sequential decision making with reinforcement learning. *Management Science*, 72(1):733–755, 2026.

Guanting Chen, Xiaocheng Li, Chunlin Sun, and Hanzhao Wang. Learning to make adherence-aware advice. *arXiv preprint arXiv:2310.00817*, 2023.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669, 2018.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, 2022.

Julien Grand-Clément and Jean Pauphilet. The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science*, 72(1):667–692, 2026.

Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning (ICML)*, pp. 9695–9728, 2022.

Matt Jensen. Rating comparisons, 2025. URL https://chessgoals.com/rating-comparison/.

Leela Chess Zero developers. Leela chess zero. https://lczero.org, 2024.

Lichess. Lichess database, 2026. URL database.lichess.org.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman ai with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1677–1687, 2020.

Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Learning models of individual behavior in chess. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Saumik Narayanan, Kassa Korley, Chien-Ju Ho, and Siddhartha Sen. Improving the strength of human-like models in chess. In *NeurIPS Workshop on Human in the Loop Learning (HiLL)*, 2022.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.

Eura Nofshin, Siddharth Swaroop, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Reinforcement learning interventions on boundedly rational human agents in frictionful tasks. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, volume 2024, pp. 1482, 2024.

Goran Peskir and Albert Shiryaev. *Optimal stopping and free-boundary problems*. Springer, 2006.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Stockfish developers. Stockfish chess engine. https://stockfishchess.org, 2026.

Eleni Straitouri, Stratis Tsirtsis, Ander Artola Velasco, and Manuel Gomez-Rodriguez. Narrowing action choices with ai improves human sequential decisions. *arXiv preprint arXiv:2510.16097*, 2025.

Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *CHI conference on human factors in computing systems extended abstracts*, 2022.

Lingguang Wang, Carlos Fernandez, and Christoph Stiller. High-level decision making for automated highway driving via behavior cloning. *IEEE Transactions on Intelligent Vehicles*, 8(1): 923–935, 2022.

Guanghui Yu and Chien-Ju Ho. Environment design for biased decision makers. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2022.

Guanghui Yu, Wei Tang, Saumik Narayanan, and Chien-Ju Ho. Encoding human behavior in information design through deep learning. *Advances in neural information processing systems (NeurIPS)*, 2023.

Yiming Zhang, Athul Paul Jacob, Vivian Lai, Daniel Fried, and Daphne Ippolito. Human-aligned chess with a bit of search. *arXiv preprint arXiv:2410.03893*, 2024.

A Model Training Details

In this section, we provide more details on the BC model used for all simulated analysis in the paper. All code is available at https://github.com/saumikn/chess_training.

A.1 Dataset Construction

All data is taken from the Lichess online database, which contains over 7.5 billion chess games played by humans. We sample from blitz games played between January 2024 and November 2025. We construct a dataset of positions taken from games played by players rated between 400 and 499. We then construct similar datasets at 500-599, etc., all the way to 2800-2899. We then uniformly sample from each of these datasets to construct our final, mixed-skill training dataset.

For all validation, testing, and base positions for our simulated results, we follow a similar dataset construction process with games played in December 2025.

A.2 Model Training

Our model is initialized with the publicly available, pre-trained Leela T82 architecture (Leela Chess Zero developers, 2024) with 768 channels, 15 blocks, and 24 attention heads. The policy head outputs an 1858-dimensional move distribution, and the value head predicts win/draw/loss probabilities.

We then train the model on the constructed dataset, with 256 million chess positions. All training code is taken exactly from the Leela open source repository, with default hyperparameters for training. The only modification we make is encoding the player rating directly into the board state (since in the standard 112x8x8 board representations, many of the layers are zeroed out because we don't encode board history). This allows us to train a single parameterized model that can represent human behavior at all skill levels between 400 and 2800.

B Exploring Human-Understandable Concepts in Our Interventions

One advantage of studying interventions in the domain of chess is that positions can be evaluated using a large number of structured, human-understandable concepts derived from classical chess evaluation functions. Modern engines such as Stockfish compute hundreds of intermediate features that capture properties of a position such as material balance, piece activity, pawn structure, and king safety. While our intervention strategy itself does not directly use these concepts, they provide a useful lens for examining the types of positions in which interventions occur.

B.1 Chess Concepts

In this section, we perform a descriptive analysis of intervention states using a subset of evaluation features derived from the Stockfish evaluation function. Specifically, we consider ten middlegame evaluation components which correspond to the highest-level intermediate terms used within the Stockfish evaluation function:

- **PieceValue:** The total material value of pieces on the board based on predefined piece weights.
- **Imbalance:** Adjustments to material evaluation based on the interaction of different piece types.
- **BishopPair:** A bonus awarded when a player possesses both bishops.
- **Pawns:** Evaluation of pawn structure features such as isolated, doubled, or supported pawns.
- **Pieces:** Evaluation of piece placement and coordination.
- **Mobility:** The number of legal squares available to pieces, reflecting piece activity and freedom of movement.
- **Threats:** Tactical pressure on opposing pieces, including attacks on undefended or weak targets.
- **PassedPawn:** Bonuses associated with pawns that have no opposing pawns blocking their path to promotion.
- **Space:** Control of central territory and safe squares in the opponent's half of the board.
- **King:** Evaluation of the king's defensive structure and exposure to potential attacks.

B.2 Methodology

We focus on the multiple-intervention setting described in Section 4.3 and analyze positions encountered under a policy tuned to produce an intervention frequency of approximately 5%. For each position encountered during simulated play, we compute the value of each Stockfish concept listed above.

To simplify analysis, we discretize each concept into three categories—low, medium, and high—based on its value in the position, based on the difference in evaluation scores for the two players. For example, PieceValue-High indicates that the player to move has more material than the opponent, PieceValue-Med indicates that both players have the same material, and PieceValue-Low indicates that the player to move has less material than the opponent.

We then compare the distribution of these categories between positions where the intervention policy overrides the human policy, and positions where no intervention occurs. For each concept and rating level, we compute the change in category frequency between intervention states and non-intervention states. Intuitively, a large difference indicates that the intervention policy correlates with positions characterized by that concept.

In Table 2, we present the top ten concepts with the largest difference between intervention and non-intervention states for each player rating.

While this analysis is purely descriptive, it suggests several preliminary patterns. For 800-rated players, the largest distribution shift is for King-High, indicating that intervention states are often

800		1200		1600		2000		2400	
Concept	Δ	**Concept**	Δ	**Concept**	Δ	**Concept**	Δ	**Concept**	Δ
King-High	0.1862	PieceValue-Med	0.0824	PieceValue-Med	0.1166	Threats-High	0.1006	PieceValue-Med	0.1116
Mobility-High	0.0848	Threats-High	0.0740	Threats-High	0.0892	PieceValue-Med	0.0692	PassedPawn-High	0.0708
PieceValue-Med	0.0780	King-High	0.0744	Mobility-Med	0.0656	Pieces-High	0.0548	Threats-High	0.0640
Threats-High	0.0734	Mobility-Med	0.0654	Pawns-Low	0.0638	King-High	0.0540	Pawns-Low	0.0576
Mobility-Med	0.0514	Pawns-Low	0.0512	Threats-Low	0.0598	Imbalance-Low	0.0456	Mobility-Med	0.0416
Pieces-High	0.0464	Imbalance-Low	0.0436	Imbalance-High	0.0528	Mobility-Med	0.0446	Threats-Low	0.0382
Imbalance-High	0.0450	Threats-Low	0.0346	Pieces-Low	0.0454	Pawns-High	0.0438	Imbalance-High	0.0340
PieceValue-High	0.0368	BishopPair-High	0.0278	Pawns-High	0.0368	Threats-Low	0.0396	Pieces-Low	0.0336
BishopPair-High	0.0314	Pieces-High	0.0216	Pieces-High	0.0316	Imbalance-High	0.0366	Pieces-High	0.0318
Space-High	0.0284	Pieces-Low	0.0214	King-Low	0.0308	BishopPair-High	0.0364	Pawns-High	0.0282

Table 2: Top concepts whose frequency differs most between intervention and non-intervention states for each rating level in the 5% intervention frequency setting. Concepts are derived from the Stockfish evaluation function and discretized into Low, Medium, and High categories. Δ denotes the absolute difference in frequency between intervention and non-intervention positions.

positions in which the opponent king is especially weak. This is consistent with the possibility that interventions at low skill levels frequently help players find decisive continuations in already promising attacking positions. At higher ratings, however, king-safety related concepts are less dominant. One possible explanation is that stronger players are already more capable of converting positions with exposed kings without assistance, so interventions become relatively more associated with other tactical or positional features.

These results should be interpreted cautiously. Because the intervention strategy itself does not explicitly use these concepts, the analysis does not imply that the model has learned explicit chess concepts or that these features causally determine intervention decisions. Instead, the analysis provides a descriptive view of the types of positions where interventions tend to occur.

Nevertheless, such concept-based analyses may provide a useful diagnostic tool for examining intervention behavior. In particular, viewing intervention decisions through the lens of established chess concepts may help identify systematic patterns in the model's behavior. More broadly, this type of analysis may offer a starting point for future work exploring how intervention strategies relate to interpretable domain knowledge, or how similar analyses could be used to diagnose weaknesses in human decision-making for educational or training purposes.