SIMPLE NEURAL TECHNIQUES FOR PLAYING GAMES OF IMPERFECT INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Developing intelligent policies for partially observable environments is challenging, as it requires reasoning about uncertainty over hidden states, planning actions under uncertain beliefs, valuing information acquisition and concealment, and finding robust policies that cannot be easily exploited. Imperfect-information games provide a natural testbed for these capabilities. In this paper, we develop simple, general techniques for efficiently training neural policies for such games, focusing on Fog-of-war chess, a game that combines the strategic complexity of chess with the challenges of imperfect information. Inspired by methods for solving imperfect-information games based on Counterfactual Regret, we adapt these ideas to reinforcement learning with large neural networks. Our approach achieves competitive performance against the state-of-the-art search techniques while doing no inference time compute or search and only requiring only a fraction of the compute used by other regret-based neural policy learning methods.

1 INTRODUCTION

Building machines that learn from their own experience is the principal and immediate goal of our field of artificial intelligence (Silver & Sutton, 2025). The challenge, however, is that the world is large and the kinds of experience possible are innumerable (Javed & Sutton, 2024). Since the world and other actors in this world will continue to evolve, the observations had and the actions that should be taken in this world will keep undergoing distribution shifts (Foerster et al., 2017); our machine must keep **continually and efficiently learning** to act in this environment (Khetarpal et al., 2022; Kumar et al., 2025). Another perspective is that large parts of the world and the policies of other agents are invisible to us; these are the source of all the non-stationarity that we experience (Khetarpal et al., 2022). It is therefore imperative that we learn to reason efficiently and robustly over imperfect information, and games of hidden information can provide us a good test bed for the same (Schmid et al., 2023; Perolat et al., 2022).

Classical reinforcement learning builds on Markov decision processes (MDPs), learning to predict long-term outcomes and improve decisions from experience (Sutton, 1988; 1990). An action induces a transition $s_t \xrightarrow{a_t} s_{t+1}$, whose probability distribution depends only on the current state and action; some states are more desirable than others, and the agent seeks actions that lead to desirable outcomes. In a partially observable Markov decision process (POMDP), the agent receives an observation o_t that provides incomplete evidence about s_t. The process given the underlying states remains Markovian, but given the observations alone does not (Kaelbling et al., 1998).

Naively, it may seem sufficient to simply apply the existing methods for playing Markovian games on observations from a game of imperfect information. However, several general techniques that achieved superhuman play in fully observable games, like AlphaZero (Silver et al., 2018) and MuZero (Schrittwieser et al., 2020), do not translate their success into games of imperfect information like poker (Brown et al., 2020a).

Following are some of the challenges faced in reasoning over imperfect information. One is that an agent must learn to retain the information from all past and present observations and use them to maintain a belief about the present underlying state (Kaelbling et al., 1998). Some hidden variables, including unobserved opponent moves, can leave multiple states and histories plausible. Efficient planning must account for their different implications under a world model, for example through

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

belief-based sampling, without treating a few trajectories as certain because they were the ones that got realized in the environment simulation (Silver & Veness, 2010; Brown et al., 2020a). Moreover, an opponent actively seeks outcomes it prefers, so we must reason jointly about our policy and its possible strategic responses (Littman, 1994).

Performance in imperfect-information games such as no-limit Texas hold'em poker (Moravčík et al., 2017; Brown & Sandholm, 2018; 2019), Hanabi (Bard et al., 2020; Lerer et al., 2020), The Resistance: Avalon (Serrino et al., 2019), Bridge bidding (Lockhart et al., 2020), StarCraft II with AlphaStar (Vinyals et al., 2019), no-press Diplomacy (Anthony et al., 2020; Gray et al., 2021; Bakhtin et al., 2021), and generals.io (Straka & Schmid, 2025; Straka et al., 2026) has served as a measure of progress in reasoning over imperfect information. Among these testbeds, Fog-of-war chess (also called dark chess) combines the demanding pattern recognition and tactical search of chess with uncertainty about the board and the opponent's knowledge. The state-of-the-art bot, Obscuro (Zhang & Sandholm, 2026), achieves superhuman play in Fog-of-war chess by combining search over imperfect information with learned evaluation of fully specified chess positions, using these state evaluations as heuristics at the search frontier.

In this paper, we study simple training techniques for learning strong neural policies and evaluation functions, for Fog-of-war chess. We start in Section 2 by placing our work in the context of known challenges and prior research on playing games of imperfect information, specifically through minimization of counterfactual regret as a way of finding Nash equilibria. Next, we translate these ideas into policy learning in Section 3. We describe the architecture of our model and a multi-stage training procedure to first learn an implicit model of the state from a sequence of observations, and then to learn a policy and an evaluation function through self-play. We show the effectiveness of this method through results in Section 4. Finally, in Section 5, we place the techniques from this paper in the context of generally reasoning over imperfect information and delineate a path that we see to this future informed by the results of our current investigations.

2 Background and Preliminaries

A two-player zero-sum partially observable stochastic game evolves through latent states $s_t \in \mathcal{S}$, with transitions $s_{t+1} \sim T(\cdot \mid s_t, a_t)$ and private observations $o_t^i = O_i(s_t)$ (Littman, 1994; Kaelbling et al., 1998). A complete history is

$$h_t = (s_0, a_0, s_1, a_1, \ldots, a_{t-1}, s_t) \in \mathcal{H}_t. \tag{1}$$

Player i observes only its private trace $\ell_i(h_t)$, comprising its observations and actions. Its infostate is the equivalence class of histories with the same trace (Kuhn, 1953; Zinkevich et al., 2007):

$$I_i(h_t) = \{h'_t \in \mathcal{H}_t : \ell_i(h'_t) = \ell_i(h_t)\}. \tag{2}$$

These classes partition $\mathcal{H}_t$: each history belongs to one infostate, while an infostate may contain many latent histories. A behavioral policy therefore maps each infostate to an action distribution, $\pi_i : I_i(h_t) \to a_t$.

2.1 Value Functions and Infostates

In MDPs, value functions summarize expected future return and support both policy improvement and search. They can be learned by temporal-difference or actor–critic methods and used as leaf evaluators in MCTS (Sutton, 1988; Konda & Tsitsiklis, 1999; Silver et al., 2017).

POMDPs need the same machinery, but values must be defined over infostates because policies act on infostates; we use CFR's counterfactual formulation (Zinkevich et al., 2007). Environment transitions, however, are defined over latent states. Let $s(h)$ be the latent state at the end of h. Given a, the simulator samples $s' \sim T(\cdot \mid s(h), a)$, receives reward $R_i(s(h), a, s')$, and forms $h' = h \circ (a, s')$. The complete history identifies a unique next infostate $I(h')$; s' alone does not, because the infostate retains the private trace. Hence $V_i^{\pi}(h, a)$ is only an intermediate simulation value: actions are selected from I, never from the hidden history h. Let $\pi_{-i}(h)$ denote the opponent-and-chance reach probability of h under π, excluding player i's contribution. With discount factor

γ and legal actions $\mathcal{A}(I)$, the value backup is

$$
\begin{aligned}
V_i^\pi(I) &= \sum_{a \in \mathcal{A}(I)} \pi_i(a \mid I) \cdot V_i^\pi(I, a), \\
V_i^\pi(I, a) &= \sum_{h \in I} \pi_{-i}(h) \cdot V_i^\pi(h, a), \\
V_i^\pi(h, a) &= \sum_{s' \in \mathcal{S}} T(s' \mid s(h), a) \cdot \Big(R_i(s(h), a, s') + \gamma V_i^\pi(I(h')) \Big).
\end{aligned}
\tag{3}
$$

A learned critic $\bar{V}_i^\pi$ approximates V_i^π and supplies the advantage $\bar{A}_i^\pi(I, a) = \bar{V}_i^\pi(I, a) - \bar{V}_i^\pi(I)$, reducing policy-gradient variance, while parameter sharing generalizes value supervision across infostates (Konda & Tsitsiklis, 1999; Schulman et al., 2016; Srinivasan et al., 2018). At each iteration t, the CFR algorithm accumulates counterfactual regret and updates the policy by regret matching (Hart & Mas-Colell, 2000; Zinkevich et al., 2007):

$$
r_i^t(I, a) = V_i^{\pi^t}(I, a) - V_i^{\pi^t}(I), \qquad R_i^T(I, a) = \sum_{t=1}^{T} r_i^t(I, a), \tag{4}
$$

$$
\pi_i^{T+1}(a \mid I) = \frac{[R_i^T(I, a)]_+}{\sum_{b \in \mathcal{A}(I)} [R_i^T(I, b)]_+} \quad \text{if} \quad \sum_b [R_i^T(I, b)]_+ > 0, \quad \text{else} \quad \frac{1}{|\mathcal{A}(I)|}. \tag{5}
$$

Values and policies must therefore co-adapt: values drive policy improvement, while each policy update changes the trajectories and returns that define those values. In self-play, both players change, creating continual distribution shift in the mappings $I \mapsto \pi_i(\cdot \mid I)$ and $I \mapsto \bar{V}_i^\pi(I)$ (Konda & Tsitsiklis, 1999; Foerster et al., 2017).

In finite two-player zero-sum games with perfect recall, vanishing regret implies convergence of the appropriately averaged policy to a Nash equilibrium (Zinkevich et al., 2007). Monte Carlo CFR uses sampled traversals, and CFR+ and predictive variants improve practical convergence (Lanctot et al., 2009; Tammelin, 2014; Farina et al., 2019; 2021).

2.2 Nash Equilibrium Search for Robust Play

Against a fixed opponent, policy gradient maximizes expected return. Against an unknown opponent, robust play instead seeks a policy that is difficult to exploit. Full observability permits search for the best move in the known state; under partial observability, the target is a least-exploitable policy characterized by Nash equilibrium.

The distinction is

$$
\max_{\pi_i} V_i(\pi_i; \pi_{-i}) \qquad \text{versus} \qquad \max_{\pi_i} \min_{\pi_{-i}} V_i(\pi_i, \pi_{-i}), \tag{6}
$$

Updating both players by self-play does not make these objectives equivalent: each player can follow a locally correct gradient against the other's current policy while their joint dynamics cycle around, or even move away from, a Nash equilibrium (Schaefer & Anandkumar, 2019; Rudolph et al., 2026). Neither ordinary perfect-information search (Brown et al., 2020a) nor naive current-policy self-play (Schaefer & Anandkumar, 2019) has a general Nash-convergence guarantee in an imperfect-information game, where the visible state does not determine an independent subgame. Nevertheless, with suitable regularization and tuning, PPO can be empirically competitive with neural fictitious-play, population, and regret-based methods (Rudolph et al., 2026).

2.3 Fictitious Play, Populations, and Evolutionary Methods

Fictitious play computes best responses to the empirical average of past opponents; NFSP approximates it with reinforcement learning and a reservoir-trained average policy (Heinrich et al., 2015; Heinrich & Silver, 2016). PSRO generalizes this idea to a restricted meta-game by adding approximate best responses to mixtures over a policy population, while evolutionary and open-ended variants preserve strategic niches in non-transitive games (Lanctot et al., 2017; Balduzzi et al., 2019).

OpenAI Five and AlphaStar implemented population memory through frozen opponents and league training (Berner et al., 2019; Vinyals et al., 2019). Replay buffers preserve examples whereas populations preserve strategic opponents; the two are complementary, and neither approximate leagues nor diversity alone guarantees Nash convergence (Silver et al., 2017; 2018; Foerster et al., 2017).

2.4 CFR-Based Search and Learning

CFR becomes an online search method by solving only the continuation relevant to the current decision while retaining counterfactual values and one strategy per infostate. DeepStack uses learned counterfactual values for continual re-solving, Libratus combines an offline blueprint with safe subgame solving, and ReBeL searches over public belief states (Moravčík et al., 2017; Brown & Sandholm, 2018; Brown et al., 2020a). When common knowledge is too large, knowledge-limited subgames restrict the search support; Obscuro combines them with growing-tree predictive CFR+ for Fog-of-war chess (Zhang & Sandholm, 2021; 2026).

Several neural methods learn CFR's regret, advantage, value, or average-strategy functions from sampled play (Waugh et al., 2015; Brown et al., 2019; Steinberger et al., 2020; McAleer et al., 2023). Student of Games combines such learning with a counterfactual value-and-policy network and Growing-Tree CFR, providing a sound search framework for both perfect- and imperfect-information games (Schmid et al., 2023). This generality is expensive: the search repeatedly expands a public tree, queries counterfactual values, and reruns CFR, while enumerating infostates at each public state can itself be prohibitive. Its large-scale comparison also used an AlphaZero baseline trained for 800,000 steps with 3,500 concurrent actors, each on a TPUv4, underscoring the distributed-compute regime in which Student of Games was evaluated (Schmid et al., 2023).

3 Method

Our method trains a transformer to consume as input a sequence of observations and output actions sampled from our policy function and a numerical evaluation of win probability.

We build the training protocol with two realizations in mind inspired by the preceding discussion:

- R1 Before reinforcement learning from self-play, we must learn from the observation history a neural representation that implicitly encodes the current infostate and its compatible histories and states. This parallels implicit neural fields (Stanley, 2007; Mildenhall et al., 2020) and Transformer-based history representations for partially observed control (Paischer et al., 2022; Dedieu et al., 2024).
- R2 On top of this implicit state we will learn policies through self-play. In doing so we must densely sample the set of possibilities in the infostate to correctly estimate policy updates, and not keep cycling between policies.

3.1 Model Architecture

Representing input observation sequences. We define an encoding of the observation at each ply which we call a POFEN (partially observable FEN) string (Edwards, 1994). POFEN retains the piece symbols and rank separators of FEN but uses one character per square, with "`.`" for visible empty squares and "`?`" for unobserved squares. Each POFEN is thus a 71-character partial board observation. Additionally, there is a 1968-move vocabulary that encompasses all possible moves on the chessboard. The context input to the model is the private history of POFEN strings interleaved with the player's own moves.

The training stream that the model uses is of the following form:
`<bos> <pos_w> POFEN_1w <move> UCI_MOVE_1w <pos_b> POFEN_1b <pos_w> POFEN_2w <move> UCI_MOVE_2w <pos_b> POFEN_2b <pos_w> ...<eos>`.

This sequence identifies the infostate of the history $I_i(h_t)$ that the player is currently in.

Policy and state-evaluation model. To represent the policy and the value function, we are using a large decoder-only transformer. It has 20 repeating attention layers and ∼530 million parameters. It

consumes the observation and move sequence and predicts the next token. This next token can be a part of the next observation that helps maintain the world model, or it can be the policy distribution over the next move. Additionally, the latent embedding used to predict the next move token is also passed to a dense value head which outputs a 128-bin evaluation. This architecture is described clearly in Appendix A of the supplementary material.

The policy function will be learned by PPO through self-play. The value function is a categorical output of 128-bins which have ordinal structure. In the process of PPO, we compute the Generalized Advantage Estimate (GAE) return as the evaluation of the state. The HL-Gauss smoothed bin distribution of the GAE return is used as a target for the cross entropy loss being used to train the value function (Farebrother et al., 2024).

$$\begin{aligned} \pi_\theta(a \mid I_i(h_t)) &\propto \mathbf{1}\{a \in \mathcal{A}(I_i(h_t))\} \exp\big(z_\theta(I_i(h_t))_a\big), \\ V_\theta(I_i(h_t)) &= \sum_{k=1}^{128} c_k \, p_\theta(k \mid I_i(h_t)). \end{aligned} \tag{7}$$

where z_θ are the policy logits and $p_\theta(k \mid \cdot)$ is the value head's probability for bin k, whose center is c_k.

For the sake of efficiency, we maintain key/value caches to retain each player's private sequence during incremental play.

History reconstruction model. A separate causal Transformer model of similar backbone architecture is trained to sample possible histories that belong to the infostate.

Its prompt first contains $I_i(h_t)$, encoded as observations and moves in the same manner as in the policy and value model and truncated at a uniformly random ply-count horizon. This is followed by a reconstruction suffix containing hypothesized joint actions and the opponent observations obtained by replaying them.

Recorded own actions are forced, while hidden opponent actions are generated. The model uses the same backbone dimensions and action vocabulary, with a 32,768-token context for the hypothesized opponent record and the reconstructed history. We sample full opponent move sequences with beam search, ranking sequences by their joint probability. We assign near-zero probability to action sequences that fail to match a known observation. This lets us effectively sample many histories $h'_t \in I_i(h_t)$. Section 4.2 shows that we sample $h'_t \in I_i(h_t)$ rather efficiently, and this would help in correctly estimating $\nabla_\theta \pi_{\theta,i}(a_t | I_i(h_t))$.

3.2 Training Protocol

Figure 1 shows how we train the model step-by-step on multiple different objectives, the motivations and specifics of which will be discussed in this section. The whole training procedure takes an estimated 250 H100 GPU hours.

Pre-training through next-observation prediction. We replay roughly 1.2 million Lichess chess games, generate fogged observations for both players, and train on one player's observations and move sequence, withholding the full board and hidden opponent moves. The pre-training loss $\mathcal{L}_{\text{pre}}$ is the standard next-token log-perplexity of this sequence (Jelinek et al., 1977; Radford et al., 2018; Brown et al., 2020b). The prediction tasks can be represented as follows:

$$O(s_1), a_1, O(s_2), a_2, \cdots, O(s_t), a_t \equiv I_i(h_t) \xrightarrow{\text{predict}} O(s_{t+1}) \tag{8}$$

Training to predict subsequent observations encourages memory of observations that are relevant to the modeling of the current state (Littman et al., 2001; Gregor et al., 2019). Predicting the moves of fully observable chess $\pi'(a_t|s_t)$ under the policies of widely varying strengths is not representative of the policy for the infostate $I_i(h_t)$ under partial information, but it can act as a reasonable warm start. Sequence models pre-trained this way on game transcripts are known to form implicit board representations (Toshniwal et al., 2022; Li et al., 2023; Karvonen, 2024); good implicit neural representations of the game state would be very useful to learn a policy on.

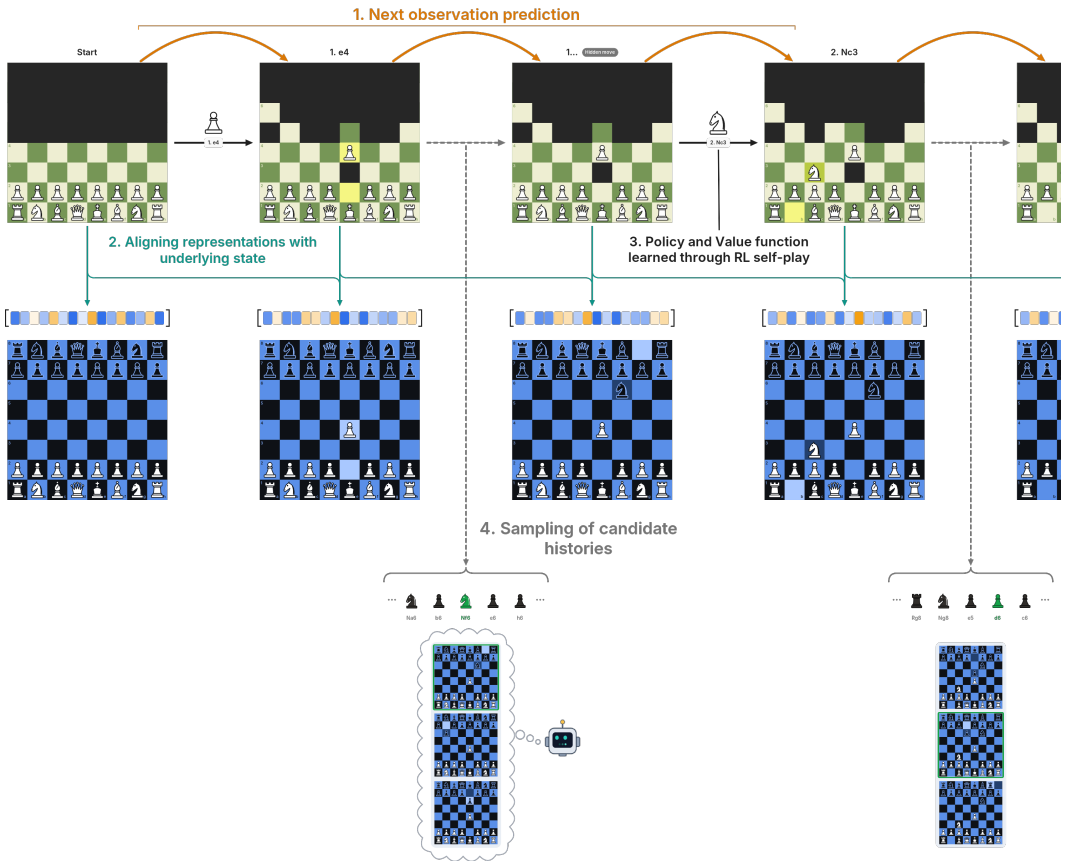


Figure 1: Overview of the method. From a player's partial-observation sequence, we (1) pre-train next-observation prediction, (2) align the learned representation with privileged state features, (3) learn policy and value heads through self-play, and (4) sample observation-consistent histories for infostate improvement.

Mid-training using privileged access to state. Searchless Chess (Ruoss et al., 2024) is a trained 270M-parameter model that is very strong at evaluating positions and playing moves in standard chess, the fully observed game underlying Fog-of-war chess. The features in the final-layer latent vector of this transformer, $\mathcal{F}_{\text{chess}}(s_t)$, which help determine the evaluation of the state, likely encode useful information about that state. To help our model build implicit representations of the infostate that capture features of the underlying state, and to bootstrap the value function, we train a projector $\mathcal{F}_\omega$ from our model's final-layer activations at the current observation's board tokens, which attend to the entire private observation history, to the latents of the Searchless Chess transformer, which has privileged access to the real board state. Additionally, we warm up the value function head by making it predict Searchless Chess's value $V_{\text{chess}}(s_t, a_t)$ of the move a_t actually played in the underlying state, compared as HL-Gauss targets generated over categorical values. The net objective, with loss weights λ_{align} and λ_{value}, thus becomes:

$$\mathcal{L}_{\text{mid}} = \mathcal{L}_{\text{pre}} + \lambda_{\text{align}} \left|\left| \mathcal{F}_\omega(I_i(h_t)) - \mathcal{F}_{\text{chess}}(s_t) \right|\right|^2 + \lambda_{\text{value}} \left|\left| V_\theta(I_i(h_t), a_t) - V_{\text{chess}}(s_t, a_t) \right|\right|. \quad (9)$$

Learning to sample candidate histories. The purpose of the history reconstruction model (Section 3.1) is to sample the possible histories in the current infostate, weighted by their opponent-and-chance reach probability: $P(h \mid I) \propto \pi_{-i}(h)$ for $h \in I$.

During the self-play procedure described next, many games are generated, and we use them to train the history reconstruction model. We take the player's observations and own moves, $I_i(h_t)$,

truncated at a uniformly random ply t, and train the model to generate the privileged opponent action trace.

Each history h_t from a real game occurs with probability $\pi_i(h_t) \cdot \pi_{-i}(h_t)$, where the reach is taken under the policies that played that game. Our own moves are forced both when training and when generating, and under perfect recall our own reach $\pi_i(h_t)$ is the same for every history in the infostate, so it cancels: the model learns to sample opponent moves according to the opponent-and-chance reach $\pi_{-i}(h_t)$. At inference we generate candidate opponent action sequences with beam search (Lowerre, 1976; Sutskever et al., 2014; Freitag & Al-Onaizan, 2017) of width 1000, ranked by the model's joint probability, and discard every sequence that fails to reproduce the observations on replay. This combines learned proposals with exact compatibility checks, as in belief sampling for partial-observation planning (Silver & Veness, 2010).

At post-training time, these sampled histories let us place identical observation sequences with different underlying histories in the same self-play training batch.

Self-play and infostate improvement. As the final step of the training procedure, once the understanding of states and infostates and a model of the transition dynamics have been baked into the model, we can begin reinforcement learning of the policy. The update follows the infostate backup of equation 3. The policy head models $\pi_i(a \mid I)$ and the value head models $V_i^\pi(I)$ explicitly, and both are trained through self-play. The action value $V_i^\pi(I, a)$ is the π_{-i}-weighted average of $V_i^\pi(h, a)$ over histories $h \in I$, and we estimate it by simulation: from a history, the environment applies the transition T and the policy plays both seats, so every completed game samples $V_i^\pi(h, a)$, bootstrapping through the value of the next infostate $V_i^\pi(I(h'))$. When the histories come from the reconstruction model (Section 3.2), the average over $h \in I$ is itself taken over π_{-i}-weighted samples; this is done only for part of training and not at all at its start, where games are played from the true history alone. The resulting advantages $V_i^\pi(I, a) - V_i^\pi(I)$ take the place of the counterfactual regrets in equation 4.

This is more faithful to CFR than Obscuro (Zhang & Sandholm, 2026), which scores the leaves of its search with an evaluation of the underlying, fully observed position, that is, with the value function of the underlying perfect-information game. That value is not $V_i^\pi(h, a)$: in an imperfect-information game the value of a history depends on what each player knows, and substituting perfect-information values introduces strategy-fusion and non-locality errors (Frank & Basin, 1998; Long et al., 2010). Because our backup always returns to the infostate $I(h')$, the value and the policy both remain defined over infostates, as the CFR equations require.

We optimize this objective with PPO (Schulman et al., 2017), using about 180 H100-hours of self-play. Each iteration freezes the behavior policy and collects 512 games: 80% current-policy self-play and 20% against a quality-weighted pool of at most 40 historical snapshots, following OpenAI Five (Berner et al., 2019). Games end naturally or after 100 plies with material adjudication. Using GAE on each player's own decision sequence (Schulman et al., 2016), we minimize

$$\mathcal{L}_{\text{post}} = \mathcal{L}_{\text{PPO-clip}} + c_v \mathcal{L}_{\text{return}} + \beta \widehat{D}_{\text{ref}} - c_H \mathcal{H}(\pi_\theta), \tag{10}$$

where $\mathcal{L}_{\text{PPO-clip}}$ is the clipped policy surrogate on the GAE advantages, $\mathcal{L}_{\text{return}}$ is the value head's cross-entropy against the HL-Gauss-encoded GAE return, $\widehat{D}_{\text{ref}}$ is a KL penalty toward an anchor policy with an adaptive coefficient β, and $\mathcal{H}$ is an entropy bonus, with loss weights c_v and c_H. The anchor starts as the supervised policy and moves to the current policy every 25 iterations, as in magnetic mirror descent (Sokota et al., 2023).

For a sampled set of histories $\widehat{I} \subset I$, we normalize the history weights to $\bar{w}_i(h \mid I) \propto \pi_{-i}(h)$ and evaluate each candidate action across the retained worlds, where $\widehat{U}_i(h, a)$ is the simulated return of playing a in h:

$$\begin{aligned} \widehat{Q}_i(I, a) &= \sum_{h \in \widehat{I}} \bar{w}_i(h \mid I)\, \widehat{U}_i(h, a), \\ \widehat{A}_i(I, a) &= \widehat{Q}_i(I, a) - \sum_{b \in \mathcal{A}(I)} \pi_i(b \mid I)\, \widehat{Q}_i(I, b). \end{aligned} \tag{11}$$

Terminal utilities are exact; nonterminal continuations use the signed value of the player to move. These estimates are conditional values, not exact CFR counterfactual values, which would also

require unnormalized reach mass. No search is run in training or at inference: we sample several rollouts and use the value head's estimates at the infostates they reach; opponent-private histories are reconstructed analogously. Approximate weights, finite support, imperfect mixing, and learned leaf values preclude inheriting exact tabular CFR guarantees.

4 RESULTS

Since our **model is the only purely learning-based technique** available to play Fog-of-war chess, we benchmark it against a baseline that does large-scale search on infostates, and against the frontier LLMs with different thinking budgets.

We post competitive win rates against both classes of models, except GPT-6 Astra Max. While our model is weaker than our implementation of Obscuro Zhang & Sandholm (2026), this is being achieved in a single forward pass with no test-time compute or search, whereas all the baseline solutions require a lot more time, compute, and multiple CPU or GPU cores, to play at their reported strengths. A public deployment is available at https://neural-pomdp.github.io/ and is playing games against humans of varied ratings with >80% win rate over >500 games.

4.1 PLAYING STRENGTH

Table 1: Head-to-head results of our policy against other models.

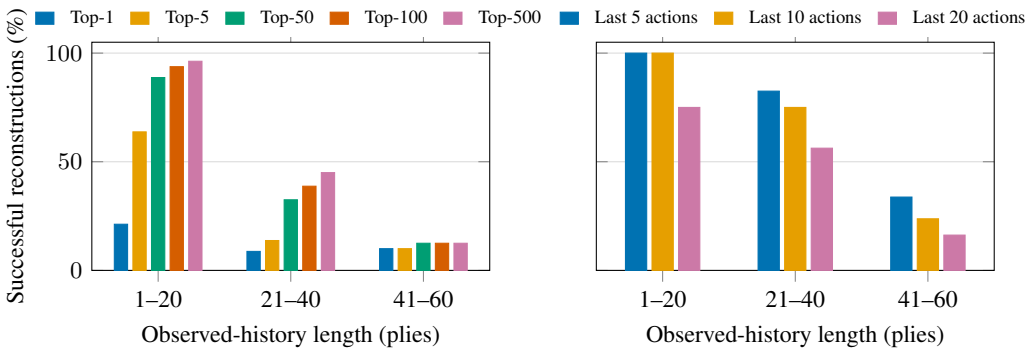
Opponent	Time per Move	Win rate	Our Relative Elo
Obscuro	0.2 s/move	31.2%	−137
Obscuro	0.5 s/move	24.1%	−199
Obscuro	1.0 s/move	14.8%	−304
GPT-6 Astra Low-thinking (OpenAI, 2026)	∼29.3 s/move	32.4%	−128
GPT-6 Astra High-thinking (OpenAI, 2026)	∼87.8 s/move	14.3%	−311
GPT-6 Astra Max-thinking (OpenAI, 2026)	∼625.8 s/move	0.0%	$-\infty$
Claude Opus 5.5 Low-Thinking (Anthropic, 2026)	∼9.4 s/move	57.9%	**+55**
Claude Opus 5.5 High-Thinking (Anthropic, 2026)	∼42.6 s/move	55.6%	**+39**

Comparison against current state-of-the-art. Obscuro (Zhang & Sandholm, 2026) is the state-of-the-art method for playing Fog-of-war chess, and the only well-studied method with strong performance. It combines GT-CFR with KLUSS to search over the game tree and uses Stockfish to evaluate the underlying game states. Since the code for Obscuro is not publicly available, we have built our own replication of the bot while remaining maximally faithful to the description of the method in the paper. We give it 3 CPU cores to search. Obscuro has stronger performance than our method, but not by a large margin. In a separate experiment that is not part of our method, adding 2s of sampling-based test-time compute improved our model by >150 Elo; no reported result uses it. Non-optimal GPU usage and lack of a great latent reasoning protocol limit this method; we discuss this in Section 5.

Comparison against frontier large language models. Language models are offering an incredibly strong baseline for any task, and merit a comparison here. In our evaluations, LLM opponents receive fixed game-specific prompts that state the rules, describe the observation and action formats, and request one move. Each opponent receives only the private information available to its player. GPT-6 Astra and Claude Opus 5.5 are evaluated at different thinking levels with unlimited thinking time per move. Table 1 reports the resulting win rates. Our model performs better than Claude. However, GPT-6 Astra is extremely strong and loses no games on max thinking. GPT-6 Astra Max also loses no games against Obscuro with 2s thinking. Understanding reasoning capabilities over partial observation and competing under low resources will be an interesting challenge.

4.2 STATE AND HISTORY RECONSTRUCTION

History reconstruction. We evaluate the History Reconstruction model and note that it is very accurate at exactly predicting the full opponent move history. The correct history, in an infoset that



(a) Exact full-history reconstruction

(b) Suffix reconstruction within beam

Figure 2: History reconstruction on four held-out games using a width-1,000 beam. (a) Frequency with which the recorded complete history occurs among the highest-probability beam candidates. (b) Frequency with which any retained history matches the recorded suffix, considering only positions with at least that many preceding actions; in the 1–20 bin, only four positions qualify for the last-20 suffix. History length and suffix length are measured in actions (plies); the persistent beam is not restarted after pruning the recorded path.

can contain over a million histories (Zhang & Sandholm, 2026), is present about **51.2%** of the time. Even when the correct history is not present, the prediction accuracy over the last few moves is pretty high, leading to effectively similar histories. Figure 2 has the details.

Next-observation prediction. Our policy+value function model also aims to maintain belief over hidden state well, to predict the next observations. We track the following metrics, which speak to our model's success at this:

- 77.7% correct prediction of where and which piece will appear when squares that were previously fogged become unfogged after our opponent's move. 78.6% on the same measure when it's our move.
- 74.6% correct predictions of which squares will be unfogged after opponent's move. 96.8% on same measure after our own move.

5 Future Work

In this paper we trained and evaluated specifically over Fog-of-war chess. Since the techniques proposed here are general, they would apply to other games of imperfect information and environments with unknown dynamics. Important challenges to this end include compressing enormous infostate spaces without merging strategically distinct states, as in generals.io (Sandholm, 2015; Straka & Schmid, 2025), and sample-efficient exploration while learning unknown dynamics, as in Private Eye under Atari100k (Kaiser et al., 2020; Ye et al., 2021).

The playing strength of our reported results comes from the neural policy alone. Test-time sampling and search over compatible histories could improve play (Silver & Veness, 2010); and latent reasoning using recurrent-depth Transformers offers another possibility (Geiping et al., 2025).

Finally, we appeal for the ability to learn strong, robust policies as a possible step towards **safer superintelligence**. Expected-return maximizers can exploit omissions in their training data distribution, e.g., if penalization of certain kinds of misaligned behavior is missing in training data, it's possible for the model to try to maximize rewards while perpetuating such behavior (Amodei et al., 2016). Optimal policy seekers also tend to try to preserve optionality and resist shutdown in certain cases (Turner et al., 2021). Policies that seek to be non-exploitable instead of maximizing value may still seek power when it improves worst-case return, but they would be more robust against missing penalties in training data mixtures. Robustness under incomplete information is therefore a possible direction that can lead to safer, more predictable models.

USE OF GENERATIVE AI

In this work, we used Generative AI tools (GPT-5.6 Sol and GPT-6 Astra, Claude Fable 5, Claude Opus 4.8 and 5.5) to help us with literature review, polishing writing, and implementation of our methods and baselines. All AI generated code has been reviewed, verified, and tested for correctness. All reported results were validated manually by the authors.

The main contents of this paper along with ideas, experiment design, and motivations come from the authors and have no origin in Generative AI. AI was used to polish the writing and fix formatting. Most citations were added by the authors, those that were suggested by AI tools have been reviewed rigorously for correctness and relevance.

We take full responsibility for the content of this paper.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565, 2016. URL https://arxiv.org/abs/1606.06565.

Thomas Anthony, Tom Eccles, Andrea Tacchetti, János Kramár, Ian Gemp, Thomas C. Hudson, Nicolas Porcel, Marc Lanctot, Julien Pérolat, Richard Everett, Roman Werpachowski, Satinder Singh, Thore Graepel, and Yoram Bachrach. Learning to Play No-Press Diplomacy with Best Response Policy Iteration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17987–18003, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d1419302db9c022ab1d48681b13d5f8b-Abstract.html.

Anthropic. Claude Opus 5.5. https://docs.anthropic.com/en/docs/about-claude/models, 2026. Model identifier: `claude-opus-5-5`; accessed 26 September 2026.

Anton Bakhtin, David Wu, Adam Lerer, and Noam Brown. No-Press Diplomacy from Scratch. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18063–18074, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/95f2b84de5660ddf45c8a34933a2e66f-Abstract.html.

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Marian Czarnecki, Julien Pérolat, Max Jaderberg, and Thore Graepel. Open-Ended Learning in Symmetric Zero-Sum Games. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 434–443. PMLR, 2019. URL https://proceedings.mlr.press/v97/balduzzi19a.html.

Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216, 2020. doi: 10.1016/j.artint.2019.103216. URL https://doi.org/10.1016/j.artint.2019.103216.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with Large Scale Deep Reinforcement Learning. arXiv preprint arXiv:1912.06680, 2019. URL https://arxiv.org/abs/1912.06680.

Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733. URL https://doi.org/10.1126/science.aao1733.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019. doi: 10.1126/science.aay2400. URL https://doi.org/10.1126/science.aay2400.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep Counterfactual Regret Minimization. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 793–802. PMLR, 2019. URL https://proceedings.mlr.press/v97/brown19b.html.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining Deep Reinforcement Learning and Search for Imperfect-Information Games. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17057–17069, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/c61f571dbd2fb949d3fe5ae1608dd48b-Abstract.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Antoine Dedieu, Wolfgang Lehrach, Guangyao Zhou, Dileep George, and Miguel Lazaro-Gredilla. Learning Cognitive Maps from Transformer Representations for Efficient Planning in Partially Observed Environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10287–10316. PMLR, 2024. URL https://proceedings.mlr.press/v235/dedieu24a.html.

Steven J. Edwards. Portable Game Notation Specification and Implementation Guide, 1994. URL https://www.saremba.de/chessgml/standards/pgn/pgn-complete.htm. Revised March 12, 1994; Section 16.1 specifies Forsyth–Edwards Notation.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taiga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop Regressing: Training Value Functions via Classification for Scalable Deep RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13049–13071. PMLR, 2024. URL https://proceedings.mlr.press/v235/farebrother24a.html.

Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-Predictive Optimistic Counterfactual Regret Minimization. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1853–1862. PMLR, 2019. URL https://proceedings.mlr.press/v97/farina19a.html.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5363–5371, 2021. doi: 10.1609/aaai.v35i6.16676. URL https://doi.org/10.1609/aaai.v35i6.16676.

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip H. S. Torr, Pushmeet Kohli, and Shimon Whiteson. Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1146–1155. PMLR, 2017. URL https://proceedings.mlr.press/v70/foerster17b.html.

Ian Frank and David Basin. Search in Games with Incomplete Information: A Case Study Using Bridge Card Play. *Artificial Intelligence*, 100(1–2):87–123, 1998. doi: 10.1016/S0004-3702(97)00082-9.

Markus Freitag and Yaser Al-Onaizan. Beam Search Strategies for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 56–60. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3207.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1380. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b01972cf31e6fa0fe29e4b8b5c2a0a1-Abstract-Conference.html.

Jonathan Gray, Adam Lerer, Anton Bakhtin, and Noam Brown. Human-Level Performance in No-Press Diplomacy via Equilibrium Search. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02923.

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aäron van den Oord. Shaping Belief States with Generative Environment Models for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2c048d74b3410237704eb7f93a10c9d7-Abstract.html.

Sergiu Hart and Andreu Mas-Colell. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68(5):1127–1150, 2000. doi: 10.1111/1468-0262.00153. URL https://doi.org/10.1111/1468-0262.00153.

Johannes Heinrich and David Silver. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. In *European Conference on Artificial Intelligence (ECAI)*, pp. 77–85, 2016. doi: 10.3233/978-1-61499-672-9-77. URL https://arxiv.org/abs/1603.01121.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious Self-Play in Extensive-Form Games. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 805–813. PMLR, 2015. URL https://proceedings.mlr.press/v37/heinrich15.html.

Khurram Javed and Richard S. Sutton. The Big World Hypothesis and its Ramifications for Artificial Intelligence. In *Finding the Frame Workshop at the Reinforcement Learning Conference*, 2024. URL https://openreview.net/forum?id=Sv7DazuCn8.

Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62 (S1):S63, 1977. doi: 10.1121/1.2016299.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłoś, Błażej Osiński, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-Based Reinforcement Learning for Atari. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xCPJHtDB.

Adam Karvonen. Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2403.15498.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards Continual Reinforcement Learning: A Review and Perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022. doi: 10.1613/jair.1.13673. URL https://doi.org/10.1613/jair.1.13673.

Vijay R. Konda and John N. Tsitsiklis. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Harold W. Kuhn. Extensive Games and the Problem of Information. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 193–216. Princeton University Press, Princeton, NJ, 1953.

Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual Learning as Computationally Constrained Reinforcement Learning. *Foundations and Trends in Machine Learning*, 18(5):913–1053, 2025. doi: 10.1561/2200000116. URL https://doi.org/10.1561/2200000116.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte Carlo Sampling for Regret Minimization in Extensive Games. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://proceedings.neurips.cc/paper/2009/hash/00411460f7c92d2124a67ea0f4cb5f85-Abstract.html.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3323fe11e9595c09af38fe67567a9394-Abstract.html.

Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving Policies via Search in Cooperative Partially Observable Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7187–7194, 2020. doi: 10.1609/aaai.v34i05.6208. URL https://doi.org/10.1609/aaai.v34i05.6208.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Michael L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the 11th International Conference on Machine Learning*, pp. 157–163, 1994. URL https://www.eecs.harvard.edu/cs286r/courses/spring06/papers/littman94markov.pdf.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive Representations of State. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1555–1561, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Edward Lockhart, Neil Burch, Nolan Bard, Sebastian Borgeaud, Tom Eccles, Lucas Smaira, and Ray Smith. Human-Agent Cooperation in Bridge Bidding. In *Cooperative AI Workshop at the Conference on Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2011.14124.

Jeffrey Long, Nathan R. Sturtevant, Michael Buro, and Timothy Furtak. Understanding the Success of Perfect Information Monte Carlo Sampling in Game Tree Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 134–140, 2010. doi: 10.1609/aaai.v24i1.7562.

Bruce T. Lowerre. *The HARPY Speech Recognition System*. PhD thesis, Carnegie Mellon University, 1976.

Stephen McAleer, Gabriele Farina, Marc Lanctot, and Tuomas Sandholm. ESCHER: Eschewing Importance Sampling in Games by Computing a History Value Function to Estimate Regret. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=35QyoZv8cKO.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, pp. 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_24. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1473_ECCV_2020_paper.php.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960. URL https://doi.org/10.1126/science.aam6960.

OpenAI. GPT-6 Astra. https://developers.openai.com/api/docs/models/gpt-6-astra, 2026. Model identifier: `gpt-6-astra`; accessed 25 September 2026.

Fabian Paischer, Thomas Adler, Vihang Patil, Angela Bitto-Nemling, Markus Holzleitner, Sebastian Lehner, Hamid Eghbal-Zadeh, and Sepp Hochreiter. History Compression via Language Models in Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17156–17185. PMLR, 2022. URL https://proceedings.mlr.press/v162/paischer22a.html.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022. doi: 10.1126/science.add4679. URL https://doi.org/10.1126/science.add4679.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Max Rudolph, Nathan Lichtlé, Sobhan Mohammadpour, Alexandre M. Bayen, J. Zico Kolter, Amy Zhang, Gabriele Farina, Eugene Vinitsky, and Samuel Sokota. Reevaluating Policy Gradient Methods for Imperfect-Information Games. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2502.08938.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, Cannada A. Lewis, Joel Veness, and Tim Genewein. Amortized Planning with Large-Scale Transformers: A Case Study on Chess. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65765–65790, 2024. doi: 10.52202/079017-2102. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/78f0db30c39c850de728c769f42fc903-Abstract-Conference.html.

Tuomas Sandholm. Abstraction for Solving Large Incomplete-Information Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. doi: 10.1609/aaai.v29i1.9757. URL https://doi.org/10.1609/aaai.v29i1.9757.

Florian Schaefer and Anima Anandkumar. Competitive Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7623–7633, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/56c51a39a7c77d8084838cc920585bd0-Abstract.html.

Martin Schmid, Matej Moravčík, Neil Burch, Rudolf Kadlec, Josh Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, G. Zacharias Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Student of Games: A unified learning algorithm for both perfect and imperfect information games. *Science Advances*, 9(46):eadg3256, 2023. doi: 10.1126/sciadv.adg3256. URL https://doi.org/10.1126/sciadv.adg3256.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. doi: 10.1038/s41586-020-03051-4. URL https://doi.org/10.1038/s41586-020-03051-4.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Jack Serrino, Max Kleiman-Weiner, David C. Parkes, and Joshua B. Tenenbaum. Finding Friend and Foe in Multi-Agent Games. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/912d2b1c7b2826caf99687388d2e8f7c-Abstract.html.

David Silver and Richard S. Sutton. Welcome to the Era of Experience, 2025. URL https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf. Preprint of a chapter in Designing an Intelligence, MIT Press.

David Silver and Joel Veness. Monte-Carlo Planning in Large POMDPs. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2164–2172, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/edfbe1afcf9246bb0d40eb4d8027d90f-Abstract.html.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017. doi: 10.1038/nature24270. URL https://doi.org/10.1038/nature24270.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://doi.org/10.1126/science.aar6404.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games. In *International Conference on Learning Representations*, 2023.

Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Perolat, Karl Tuyls, Remi Munos, and Michael Bowling. Actor-Critic Policy Optimization in Partially Observable Multiagent Environments. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/e22dd5dabde45eda5a1a67772c8e25dd-Abstract.html.

Kenneth O. Stanley. Compositional Pattern Producing Networks: A Novel Abstraction of Development. *Genetic Programming and Evolvable Machines*, 8(2):131–162, 2007. doi: 10.1007/s10710-007-9028-8. URL https://doi.org/10.1007/s10710-007-9028-8.

Eric Steinberger, Adam Lerer, and Noam Brown. DREAM: Deep Regret minimization with Advantage baselines and Model-free learning. arXiv preprint arXiv:2006.10410, 2020. URL https://arxiv.org/abs/2006.10410.

Matej Straka and Martin Schmid. Artificial Generals Intelligence: Mastering Generals.io with Reinforcement Learning. arXiv preprint arXiv:2507.06825, 2025. URL https://arxiv.org/abs/2507.06825.

Matej Straka, Viliam Lisý, and Martin Schmid. Superhuman AI for Generals.io Using Self-Play Reinforcement Learning. arXiv preprint arXiv:2606.23348, 2026. URL https://arxiv.org/abs/2606.23348.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27, pp. 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

Richard S. Sutton. Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224, 1990. doi: 10.1016/B978-1-55860-141-3.50030-4. URL https://doi.org/10.1016/B978-1-55860-141-3.50030-4.

Oskari Tammelin. Solving Large Imperfect Information Games Using CFR+. arXiv preprint arXiv:1407.5042, 2014. URL https://arxiv.org/abs/1407.5042.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a Testbed for Language Model State Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11385–11393, 2022. doi: 10.1609/aaai.v36i10.21390.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal Policies Tend To Seek Power. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23063–23074, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. doi: 10.1038/s41586-019-1724-z. URL https://www.nature.com/articles/s41586-019-1724-z.

Kevin Waugh, Dustin Morrill, J. Andrew Bagnell, and Michael Bowling. Solving Games with Functional Regret Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. doi: 10.1609/aaai.v29i1.9445. URL https://doi.org/10.1609/aaai.v29i1.9445.

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering Atari Games with Limited Data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25476–25488, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d5eca8dc3820cad9fe56a3bafda65ca1-Abstract.html.

Brian Zhang and Tuomas Sandholm. Subgame solving without common knowledge. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23993–24004, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c96c08f8bb7960e11a1239352a479053-Abstract.html.

Brian Hu Zhang and Tuomas Sandholm. General search techniques without common knowledge for imperfect-information games, and application to superhuman Fog of War chess. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=afaakBqkvb.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret Minimization in Games with Incomplete Information. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/hash/08d98638c6fcd194a4b1e6992063e944-Abstract.html.