LEARNING TO BARGAIN THROUGH AN ALLOCATION-SENSITIVE VALUE BRIDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning to bargain within a larger sequential decision problem requires agents to identify which exchanges to make and which proposals others will accept. When the value of an exchange is endogenous to the surrounding task, agents face sparse and high variance reward signals during bargaining. We introduce a hierarchical reinforcement-learning framework that addresses this by learning an allocation-sensitive value bridge that provides immediate post-bargaining feedback without prescribing agent behavior or relying on negotiation demonstrations. We test this framework in a four-player variant of the board game Settlers of Catan with two independently seeded self-play league lineages. Our experiments demonstrate that agents trained with the value bridge win more often against common opponents, reach agreement more often, and rely more on negotiated player exchange. This performance improvement is attributable to gains in bargainer ability. Ablations show that without the value bridge the bargainer loses selectivity, performs no better than a random bargainer, and the complete agent wins well below par. These results show that economically recognizable bargaining behavior can emerge from task-oriented self-play using a learned interface between negotiation and long-horizon control.

1 INTRODUCTION

Bargaining is often embedded within a larger sequential decision problem. In such settings, an agent must determine not only which proposals others will accept, but also which exchanges will improve its downstream prospects. This creates a difficult credit-assignment problem: the value of an exchange depends on the surrounding state and policy, while its consequences may become apparent only after many subsequent decisions. Bargaining from delayed task reward alone can provide a sparse, high-variance signal, while assigning fixed values to exchanged assets ignores their state-dependent strategic value. The central challenge is therefore to provide bargaining-specific feedback without prescribing which exchanges should be made.

We introduce an allocation-sensitive *value bridge* for learning bargaining via hierarchical reinforcement learning. We represent the environment as two linked stochastic games: a macro game governing the surrounding control problem and a state-conditioned bargaining game governing proposed reallocations. A macro policy decides when to initiate bargaining, and a specialized bargaining policy constructs proposals and responds to other agents. To connect these policies, we train an auxiliary critic to predict macro-game returns from a restricted representation of the current resource allocation and strategic context; this value-bridge does not impose fixed asset prices or prescribe bargaining behavior. When bargaining changes the allocation, the corresponding change in the auxiliary critic's estimate supplies immediate feedback to the bargaining policy.

We study the framework in a four-player implementation of the popular board game *Settlers of Catan*, where resource exchanges influence a long sequence of stochastic, strategic decisions. Across two independent self-play lineages, the complete hierarchy improves its win rate against common opponents during training. Policy-swap experiments localize much of this improvement to the learned bargaining policy. Training also produces more selective proposals, fewer but more successful counteroffers, and greater effective use of bargaining. Targeted ablations further indicate that removed or poorly aligned bridge signals fail to reproduce the performance of the complete system. Together, these results provide evidence that economically recognizable bargaining behav-

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.

ior can emerge from task-oriented self-play when a learned value interface connects negotiation to long-horizon control.

Our contributions are threefold:

1. We formulate task-embedded bargaining as a state-conditioned stochastic game linked to a surrounding control problem through bargaining initiation, implemented reallocations, and macro continuation value.
2. We introduce an allocation-sensitive value bridge that converts changes in a learned, policy-aligned continuation-value surrogate into immediate feedback for a specialized bargaining policy.
3. We evaluate the hierarchy using common-opponent comparisons, policy swaps, behavioral measurements, and targeted ablations, finding improved task performance and increasingly selective bargaining across two self-play lineages.

2 Related Work

Learning to bargain. Learned bargaining has been studied in structured allocation problems, dialogue systems, and multi-agent environments. Prior approaches learn from negotiation corpora, self-play, rollout-based planning, or explicit negotiation outcomes (Lewis et al., 2017; He et al., 2018; Cao et al., 2018; Foerster et al., 2018). More recent work uses pretrained or fine-tuned language models to negotiate through natural-language interaction (Fu et al., 2023; Abdelnabi et al., 2024; Cao et al., 2025). These settings primarily study communication, reasoning, or agreement under a specified bargaining objective. We instead study a structured bargaining protocol inspired by (Rubinstein, 1982), (Binmore et al., 1986), (Osborne & Rubinstein, 1990), and (Krishna & Serrano, 1996) embedded in a stochastic game, where preferences over exchanges are induced by the surrounding task and change with the strategic state. Our agents receive neither negotiation demonstrations nor fixed utilities for the exchanged resources.

Hierarchical reinforcement learning and credit assignment. Hierarchical reinforcement learning decomposes long-horizon problems into temporally or functionally distinct controllers (Sutton et al., 1999; Kulkarni et al., 2016; Vezhnevets et al., 2017). Our hierarchy is specified by the environment's institutional structure: a macro policy controls the surrounding game, while a bargaining policy acts within bargaining instances. The principal challenge is not discovering this decomposition, but communicating long-horizon task value to the bargaining policy.

Our value bridge is related to auxiliary value prediction, reward shaping, and methods for assigning delayed returns to earlier decisions (Jaderberg et al., 2017; Ng et al., 1999; Arjona-Medina et al., 2019). Classical potential-based shaping preserves optimal policies when rewards are transformed using an appropriately discounted difference of a fixed potential (Ng et al., 1999). Our bridge instead uses the change in a concurrently learned, allocation-sensitive continuation-value estimate to train a separate policy. Because the bargaining policy discounts bargaining rewards differently and the learned potential changes during training, the standard policy-invariance guarantee does not apply. We treat the bridge as an approximate, policy-aligned credit-assignment mechanism and evaluate its effects through task performance, policy swaps, and targeted ablations.

Catan and task-embedded bargaining. Early work on Catan studied planning, reinforcement learning, and handcrafted or learned policies for its large, stochastic action space (Thomas, 2003; Szita et al., 2010; Xenou et al., 2019; Gendre & Kaneko, 2020). These systems either disable in-game trading, fix it to a hand-written heuristic, or fold it into a single flat action space alongside every other game action. Closest to our setting, Cuayáhuitl et al. (2015) and Keizer et al. (2017) trained strategic dialogue policies to propose, accept, reject, and counter resource exchanges in Catan. This work established that bargaining behavior can be learned from interaction. Our focus is complementary: we jointly train a specialized bargaining policy and a policy for the surrounding sequential task, and derive bargaining feedback from a learned estimate of the reallocations' downstream task value. The contribution is therefore not the introduction of learned trading in Catan, but a learned value interface between bargaining and long-horizon control.

3 Problem Setting and Hierarchical Formulation

We consider environments in which a structured bargaining process is embedded within a larger sequential decision problem. In this situation, the full decision problem contains two qualitatively different classes of actions, those which alter the environment and those which specifically construct and evaluate proposed reallocations of assets.

We model this structure using two linked stochastic games: a macro game G_M, which governs the outer control problem, and a state-conditioned bargaining game $G_B(\cdot)$, which governs negotiation over transferable assets. A macro action initiates the bargaining game, and the terminal outcome of bargaining determines the state from which the macro game resumes.

3.1 The Macro Game

We represent the outer, macro game as a tuple:

$$G_M = (\mathcal{N}, \mathcal{S}_M, \{\mathcal{A}_{M,i}\}_{i\in\mathcal{N}}, p_M, \{r_{M,i}\}_{i\in\mathcal{N}}, \gamma_M)\,.$$

The set of agents is $\mathcal{N} = \{1, \ldots, n\}$. The state space in the macro game is $\mathcal{S}_M$. The set $\mathcal{A}_{M,i}(s)$ are the feasible actions of agent i in state s. Let $a_t = (a_{1,t}, \ldots, a_{n,t}) \in \prod_{i\in\mathcal{N}} \mathcal{A}_{M,i}(s_t)$ denote the joint macro action, where t indexes macro game time;[1] The macro game environmental dynamics are governed by a transition kernel so that $s_{t+1} \sim p_M(\cdot|s_t, a_t)$. The macro game reward function for agent i is $r_{M,i}(s_t, a_t, s_{t+1})$. Finally, the discount factor in the macro game is $\gamma_M \in [0, 1)$.

At each time t, agent i holds an allocation of K transferable asset classes, say $\alpha_i^{(t)} = \left(\alpha_{i,1}^{(t)}, \ldots, \alpha_{i,K}^{(t)}\right) \in \mathbb{R}^K_{\geq 0}$. Let $\alpha^{(t)} = \left(\alpha_1^{(t)}, \ldots, \alpha_n^{(t)}\right)$ denote the joint allocation. We write $s_t^\alpha \in \mathcal{S}_M$ for a macro state with allocation α. The superscript identifies the allocation component of the state: if the nonallocation components remain fixed while the allocation changes from α to α', we denote the resulting state by $s_t^{\alpha'}$.[2]

Each agent samples actions in the macro game according to a policy so that $a_{i,t} \sim \pi_{M,i}(\cdot \mid s_t^\alpha)$. The environment then transitions to a new state and produces a reward for i.

3.2 The State Conditioned Bargaining Game

The macro action space contains an action that initiates bargaining. Selecting this action temporarily transfers control from the macro game to a bargaining game conditioned on the current macro state. Thus, for each macro state s_t^α at which bargaining is initiated, we define a state-conditioned bargaining game, also represented as a tuple:

$$G_B(s_t^\alpha) = (\mathcal{N}, \mathcal{S}_B(s_t^\alpha), \{\mathcal{A}_{B,i}(\cdot|s_t^\alpha)\}_{i\in\mathcal{N}}, p_B, \{r_{B,i}\}_{i\in\mathcal{N}}, \gamma_B)\,.$$

Here, we focus on the general requirements of this bargaining game; Section 4.1 specifies the particular protocol used in our experiments.

The set of agents $\mathcal{N}$ is shared with the macro game. The set of states in the bargaining game is $\mathcal{S}_B(s_t^\alpha)$. The initiating macro state supplies the strategic context for the bargaining game, including the incumbent allocation and all nonallocation state information relevant to the continuation of the macro game. A state of the bargaining game is written $s_{B,\tau} = (s_t^\alpha, \alpha_{B,\tau}, h_\tau)$, where $\alpha_{B,\tau}$ is the current allocation during bargaining, h_τ records the current protocol phase, and τ indexes bargaining time. Depending on the protocol, h_τ may record current proposer, pending proposals and their affected counterparties, responses, proposal counts, and/or remaining event budget. The initiating macro context s_t^α remains fixed during the bargaining episode, while $\alpha_{B,\tau}$ and h_τ evolve according to the agents' actions during bargaining.

[1] In turn-based environments, inactive agents may be assigned a null action.

[2] We assume perfect monitoring of the current allocation and all bargaining-relevant public information. The stochastic-game state determines environmental transitions, but policies condition on agent information states containing the publicly monitored allocation, private observations, and any belief-state features used by the agent. To avoid additional notation, we suppress this distinction in the general presentation.

Let $\mathcal{F}(s_t^\alpha) \subseteq \left(\mathbb{R}^K_{\geq 0}\right)^n$ denote the set of allocations that can feasibly result from bargaining initiated at s_t^α.[3]

Let $q(s_{B,\tau})$ denote the active proposer and let $C(s_{B,\tau}) \subseteq \mathcal{N}$ denote the set of agents whose acceptance is required for the current proposal to become the current allocation.[4] Define the role-dependent bargaining action set as

$$\mathcal{A}_{B,i}(s_{B,\tau}|s_t^\alpha) = \begin{cases} \mathcal{F}_i(s_t^\alpha), & i = q(s_{B,\tau}), \\ \{\text{ACCEPT}, \text{REJECT}\}, & i \in C(s_{B,\tau}), \\ \{\varnothing\}, & \text{otherwise}, \end{cases}$$

where $\mathcal{F}_i(s_t^\alpha) \subseteq F(s_t^\alpha)$ is the set of proposals available to proposer i. Bargaining states evolve according to p_B, which captures the order of proposals and responses, the implementation of accepted reallocations, and the protocol's termination rules. The discount factor in the bargaining game is γ_B. We discuss bargaining game rewards, $r_{B,i}$, below.

Each agent samples actions in the state conditioned bargaining game according to a policy so that $a_{B,i,\tau} \sim \pi_{B,i}(\cdot \mid s_{B,\tau})$. The bargaining game terminates at stopping time $T_B = \inf\{\tau | s_{B,\tau} \in \mathcal{S}_B^{\text{term}}(s_t^\alpha)\}$, where $\mathcal{S}_B^{\text{term}}(s_t^\alpha)$ denotes the set of terminal states of the bargaining game initiated at macro state s_t^α. In general, we will require that $T_B < \infty$ almost surely.

3.3 Linking the Two Games

The macro and bargaining games are linked by initiation and return maps, denoted as $I_i(s_t^\alpha)$ and $R(s_{B,T_B})$, respectively. In general, we set

$$I_i(s_t^\alpha) = (s_t^\alpha, \alpha_{B,0}, h_0)$$
$$R(s_{B,T_B}) = s_{t+1}^{\alpha_{B,T_B}}$$

where $\alpha_{B,0} = \alpha$.[5] For each possible initiating agent i, $I_i(\cdot)$ provides the bargaining game with an initial state. Similarly, $R(\cdot)$ updates the state in the macro game with the (potentially) new allocation at bargaining termination, along with any additional information relevant to the state dimensions in the macro game.

Thus, given a measurable subset of macro game states, $S \subseteq \mathcal{S}_M$, the bargaining game induces a state-dependent macro transition kernel under the bargaining-policy profile π_B:

$$K_{\pi_B}(S \mid s_t^\alpha) = \Pr_{\pi_B, p_B}\left(R(s_{B,T_B}) \in S \mid s_{B,0} = I_i(s_t^\alpha)\right).$$

The state at which bargaining begins, along with the bargaining policies, determine the distribution over allocations from which the macro game resumes. For ordinary macro actions, state transitions are governed directly by p_M. For the initiate-bargaining action, K_{π_B} supplies the corresponding state-transition distribution.

3.4 An Allocation-Sensitive Value Bridge

We treat the value of asset exchanges during bargaining as endogenous to the surrounding control problem.[6] The same transfer may have different values in different macro states or under different macro policies.

[3] Dependence on the macro state allows for the preexisting allocation to constrain what future allocations are possible. In general, the environmental rules will determine which allocations are feasible. Reasonable conditions include things like no-debt, i.e. $\alpha'_{i,k} \geq 0$ for every $i \in \mathcal{N}$ and $k \in \{1, \ldots, K\}$, and asset-by-asset conservation, $\sum_{i\in\mathcal{N}} \alpha'_{i,k} = \sum_{i\in\mathcal{N}} \alpha_{i,k}$ for every $k \in \{1, \ldots, K\}$.

[4] For example, with simultaneous unanimous voting we may define the required voter set as $C(s_{B,\tau}) = \{j \in \mathcal{N} \backslash \{q(s_{B,\tau})\} | \alpha'_j \neq \alpha_j\}$.

[5] Note that the returned state is indexed forward by one time period; this is effectively an assumption that bargaining is instantaneous from the perspective of the macro game, i.e. that the initiate bargaining action is not more temporally costly than any other action.

[6] This is motivated by the difficulty of formulating general utility functions for asset holding; even a simple rule like 'more is better' is not universally applicable, and indeed is not necessarily true in the Catan environment we consider below.

To deal with this, we use the macro game state-value difference before and after bargaining as a *value bridge*: macro-game returns train a macro game state value function, and changes in that provide rewards during bargaining. Specifically, for a fixed policy profile $\pi = \{\pi_{M,1}, \ldots, \pi_{M,n}, \pi_{B,1}, \ldots, \pi_{B,n}\}$, define agent i's macro state value as

$$V^{\pi}_{M,i}(s) = \mathbb{E}_{\pi, p_M, p_B}\left[\sum_{\ell=0}^{T-t-1} \gamma_M^{\ell} r_{M,i}(s_{t+\ell}, a_{t+\ell}, s_{t+\ell+1}) \middle| s_t = s\right].$$

If bargaining changes the allocation from α to α' then, relative to disagreement, the change in state value of the proposed reallocation is

$$A^{\pi}_{B,i}\left(s_t^{\alpha}, s_t^{\alpha'}\right) = V^{\pi}_{M,i}\left(s_t^{\alpha'}\right) - V^{\pi}_{M,i}\left(s_t^{\alpha}\right).$$

This is the ideal continuation-value contrast associated with moving from one allocation to another. We seek to provide this quantity to the bargaining policy as a reward.

Learning the value of asset exchange via its impact on the macro game is a natural choice but does imply that the bargaining policy faces a credit-assignment problem. A reallocation may influence environmental rewards only after many subsequent macro decisions, a potentially sparse and high-variance learning signal. In high-dimensional state representations, a bargain might change only a small subset of a critic's inputs. Without management of how the consequences of those changes are represented, the full-state value difference may be too small or noisy to provide a useful learning signal.

We therefore construct a surrogate: an allocation-sensitive auxiliary state value function with a restricted state representation denoted $C^{\psi}_{\alpha,i}\left(\psi_i(s_t^{\alpha})\right)$, where $\psi_i(\cdot)$ projects the macro state onto a lower-dimensional set of features selected to represent the strategic consequences of asset ownership for agent i.

If bargaining is initiated at s_t^{α} and terminates with an accepted allocation α' we can use the auxiliary state value function as a learned surrogate for the ideal continuation-value contrast. Thus, define

$$\Delta^{\pi}_{\alpha,i}\left(s_t^{\alpha}, s_t^{\alpha'}\right) \equiv C^{\pi}_{\alpha,i}\left(\psi_i(s_t^{\alpha'})\right) - C^{\pi}_{\alpha,i}\left(\psi_i(s_t^{\alpha})\right).$$

This is our value bridge. Given this, we specify the following contribution to the bargaining game rewards:

$$r_{B,i}(s_{B,\tau}, a_{B,\tau}, s_{B,\tau+1}) = \mathbf{1}\{s_{B,\tau+1} \in S_B^{\text{term}}(s_t^{\alpha})\}\Delta^{\pi}_{\alpha,i}\left(s_t^{\alpha}, R(s_{B,\tau+1})\right).$$

Because this construction depends only on the projected state, $\psi_i(s)$, its difference should be interpreted as an approximate, macro policy-aligned shaping signal rather than the full causal effect of a reallocation.[7]

The macro policy acts in the outer stochastic game and initiates the bargaining game when appropriate. The bargaining policies propose and evaluate reallocations. When an allocation is implemented, the auxiliary resource state value function converts the resulting change in policy-relevant state into an immediate post-bargaining learning signal, and the terminal bargaining state determines the macro state from which play resumes. We depict this architecture in Figure 2 in Appendix A.2.

3.5 Reward Components

We provide the state value function, $V^{\pi}_{M,i}(\cdot)$, the full returns derived from $r_{M,i}(\cdot)$ during training. The allocation-sensitive auxiliary state value function used in the value bridge is trained on a separate reward channel containing $r_{M,i}$ plus additional potential-based reward shaping. For a stored macro transition, let $\Phi_{\alpha,i}(\cdot)$ denote an asset-use potential evaluated from agent i's asset allocation related to macro game actions that become possible or impossible depending on asset allocation. The allocation-sensitive auxiliary state value function reward is

$$r^{(\alpha)}_{M,i}(s_t, a_t, s_{t+1}) = r_{M,i}(s_t, a_t, s_{t+1}) + \beta[\gamma_M \Phi_{\alpha,i}(s_{t+1}) - \Phi_{\alpha,i}(s_t)].$$

[7] This also differs from classical potential-based reward shaping. In particular, this value bridge is learned concurrently rather than held fixed. We consequently do not claim the standard policy-invariance guarantee associated with potential-based shaping. The empirical question is instead whether the allocation-sensitive bridge produces policies whose reallocations improve independently measured macro-game outcomes.

This does not impose fixed utilities or exogenous prices on the transferable assets. The restricted input makes changes in the state dimensions controlled by the bargaining policy salient to the bargaining reward. Its task-return component is endogenous to the macro game, while the shaping potential supplies an inductive bias toward allocations that enable macro actions.

The auxiliary return, regression loss, and target construction are given in Appendix A.2.

4 Environment and Training

We instantiate the task–bargaining framework in the popular board game Settlers of Catan, where resource exchanges have state-dependent value and require another player's consent.

4.1 Catan Environment and Bargaining Protocol

The environment is a four-player implementation of Settlers of Catan with five asset classes and a ten-victory-point game completion threshold. Asset allocations are publicly observable while other cards held by agents are private.

The macro policy controls all non-bargaining game actions including the initiation of bargaining. At most one bargaining game may be initiated per player turn; these bargaining games have a ten-event limit and the initiating agent is the proposer. Allocation offers specify a target for the proposal (an agent or the heuristic Catan bank that trades assets at fixed rates) and separate give and receive bundles; agents may accept the proposal or counteroffer to the proposer. The first accepted exchange executes and ends the bargaining game. Other termination conditions include voluntary termination and exhaustion of the event budget.

4.2 Learning Architecture

We train agents with PPO (Schulman et al., 2017) and generalized advantage estimation (Schulman et al., 2016). The macro policy and primary critic encode the agent's information state, which combines public game information, the agent's private information, and belief-state features for hidden opponent information. The allocation-sensitive auxiliary state value function used in the value bridge uses a separate low-dimensional representation emphasizing resource holdings, public player status, and construction opportunities while omitting board geometry. The bargaining network encodes participants, pending and recent offers, and strategic context, and produces a structured proposal through conditionally masked categorical action heads. The macro policy, primary critic, and allocation-sensitive critic have separate parameters, whereas the bargaining actor and critic share an encoder. Architectural dimensions and optimization settings are reported in Appendix A.1 and A.3.

4.3 Reward Channels and Auxiliary Critic Training

The macro game reward function rewards only changes in victory points; crossing the ten-point threshold gives a reward of 10 in place of the ordinary point increment. The macro policy update uses advantages constructed from this macro game reward.

An executed exchange yields a normalized and clipped value bridge reward based on the allocation-sensitive auxiliary critic's pre/post-bargaining difference; unsuccessful bargaining games and intermediate bargaining actions receive no reward. Full information on the reward structure is in Appendix A.4.

4.4 Self-Play League Training

We train via self-play using a league inspired by Vinyals et al. (2019) containing frozen historical policies and specialized exploiters. Opponents are sampled using prioritized fictitious self-play, and their macro and bargaining components may be drawn independently to increase strategic diversity. Full league composition, snapshot admission, and sampling procedures appear in Appendix A.6.

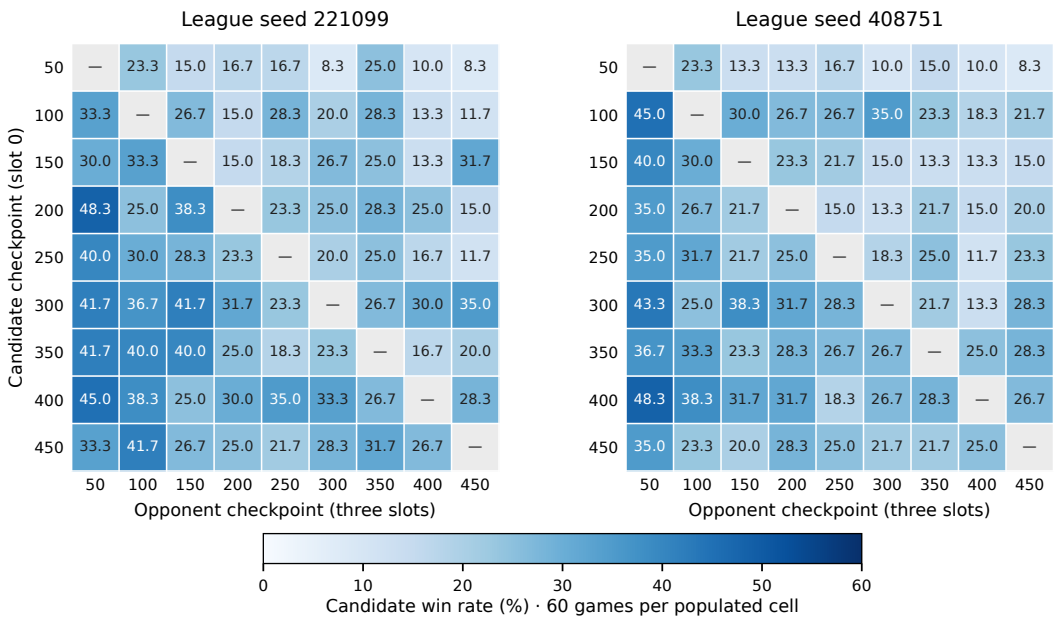

Figure 1: Completed PFSP league tournaments: 72 ordered checkpoint matchups per lineage and 8,640 games in total. Each populated cell reports the candidate's win rate in percent over 60 games against three copies of the column checkpoint. Rows and columns both denote complete hierarchies.

5 Results and Discussion

We provide results from two independently seeded training lineages that share the task, architecture, and reward-bridge recipe. We first ask whether complete hierarchies improve during league training. We then use policy swaps to identify whether any performance gain is attributable to the macro policy or bargainer, characterize how bargaining behavior changes, and evaluate whether the allocation-sensitive value bridge is necessary for learning. Secondary analyses of initial placement, game duration, reward magnitudes, and early development failures appear in Appendices A.10, A.11, A.12, and A.13.

5.1 Performance of the Complete Hierarchy

We compare the update-50 and update-450 agents against the same intermediate opponents within each league lineage. In lineage 221099, win rate increases from 16.4% to 28.8%, a paired gain of 12.4 percentage points with 95% confidence interval of [5.5,19.5]. In lineage 408751, win rate increases from 14.5% to 23.6%, a gain of 9.0 points with interval [2.6,15.5]. Thus, both realized training lineages improve against their respective common opponent sets. See Table 4 in Appendix A.9.

The complete checkpoint tournaments show that performance generally rises through updates 300–400 before plateauing; full matchup matrices appear in Figure 1. These comparisons establish within-lineage progress, but do not imply monotonic improvement at every checkpoint or measure generalization beyond the trained opponent populations.

5.2 Attributing Improvement to the Bargainer

The checkpoint tournaments vary the macro policy and bargainer together. To identify which component accounts for the observed improvement, we cross the update-50 and update-400 policies from each lineage and add a random-bargainer baseline. This produces five conditions: $(C1 = (M_{50}, B_{50}))$, $(C2 = (M_{50}, B_{400}))$, $(C3 = (M_{400}, B_{50}))$, $(C4 = (M_{400}, B_{400}))$, and $(C5 = (M_{400}, B_{\text{random}}))$. Each condition plays 256 paired games against both a same-lineage update-400 opponent panel and a mixed-strength cross-lineage panel.

Table 1: Paired policy-swap effects: within-episode win-rate differences of the evaluated seat in percentage points with 95% paired bootstrap intervals (256 pairs per cell). Every bargainer contrast excludes zero; every macro policy contrast includes it.

Contrast		221099, Panel A	221099, Panel B	408751, Panel A	408751, Panel B
learned vs. random bargainer	C4 − C5	$+21.1$ $[16.0, 26.6]$	$+23.4$ $[18.0, 28.9]$	$+19.5$ $[14.5, 24.6]$	$+25.8$ $[20.3, 31.2]$
bargainer U50→U400, late macro policy	C4 − C3	$+11.7$ $[6.6, 16.8]$	$+12.9$ $[7.0, 18.8]$	$+8.2$ $[2.7, 13.7]$	$+14.8$ $[9.0, 20.7]$
bargainer U50→U400, early macro policy	C2 − C1	$+9.0$ $[3.5, 14.8]$	$+8.6$ $[3.1, 14.1]$	$+6.3$ $[0.8, 11.7]$	$+10.2$ $[5.1, 15.6]$
macro policy U50→U400, early bargainer	C3 − C1	-4.3 $[-9.4, 0.8]$	-0.8 $[-5.9, 3.9]$	-0.8 $[-5.9, 4.3]$	-5.5 $[-11.3, 0.4]$
macro policy U50→U400, late bargainer	C4 − C2	-1.6 $[-7.8, 5.1]$	$+3.5$ $[-2.3, 9.4]$	$+1.2$ $[-5.1, 7.8]$	-0.8 $[-7.4, 5.9]$
interaction	(C4−C3)−(C2−C1)	$+2.7$ $[-5.1, 10.2]$	$+4.3$ $[-2.3, 11.3]$	$+2.0$ $[-5.9, 9.8]$	$+4.7$ $[-2.3, 12.5]$

Table 1 reports the paired contrasts. Across both lineages and opponent panels, replacing the early bargainer with the late bargainer increases win rate by 6.3 to 14.8 percentage points, with every paired 95% confidence interval excluding zero. Replacing the late bargainer with a random one reduces win rate by 19.5 to 25.8 points. By contrast, replacing the early macro policy with the late one changes win rate by -5.5 to 3.5 points, and every corresponding interval includes zero; the estimated policy interactions are likewise small and uncertain. Thus, within the resolution of this experiment, the evaluated improvement between updates 50 and 400 is attributable to the bargainer, while the later macro policy provides no detectable additional win-rate advantage. Raw win-rates are available in Appendix A.9.

5.3 Emergent Bargaining Behavior

We compare agent bargaining behavior at updates 50 and 450 against common opponents. In the two league lineages, the fraction of initiated sessions that execute an exchange rises from 58.8% to 69.3% and from 57.1% to 72.1%, even as initiations per game decline slightly. The later policies therefore convert bargaining attempts into agreements more effectively rather than simply initiating negotiation more often.

Bargaining also shifts toward strategic counterparties. Player exchanges per game increase from 4.4 to 8.5 and from 4.4 to 9.2, while bank exchanges remain approximately constant. Consequently, the bank's share of executed exchanges falls from approximately 60% to 42%. Counteroffers become less frequent but more likely to be accepted: counteroffers per game fall from 19.8 to 14.5 and from 23.4 to 16.1, while acceptance increases from 7.6% to 9.0% and from 7.0% to 8.4%. Paired policy-swap games reproduce these behavioral differences when the macro policy is held fixed, attributing them to the bargainer rather than to changes in macro control.

These descriptive patterns are economically recognizable but are not prescribed by the reward design: neither the shaping potential nor the completion bonus specifies an agreement rate, counteroffer frequency, or division of trade between the bank and other players. Their joint emergence through self-play, together with the policy-swap results, indicates that the bargainer learns a more selective and effective use of bargaining.

5.4 Value-Bridge Ablation

We evaluate whether useful bargaining depends on the allocation-sensitive value bridge in a separate six-victory-point experiment. Three value bridge-trained runs serve as controls. Two single-seed ablations replace the bridge with either the primary macro critic's pre/post-bargaining value difference or a completion reward with no learned value difference. All conditions otherwise use the same environment, architecture, and training procedure.

The value bridge-trained controls remain selective: their training agreement rates are 46%–49% by update 50 and rise to 57%–62% by update 150. By contrast, both ablations converge toward near-universal agreement. The no-bridge agent executes almost every initiated bargain from update 20 onward, while the full-state agent reaches agreement in 94%–98% of attempts by updates 40–80. Without an allocation-sensitive signal, the completion incentive rewards agreement without distinguishing beneficial from harmful exchanges and the bargaining policy collapses toward near-universal acceptance. Earlier unmatched development runs exhibited the opposite collapse toward near-zero agreement under different residual incentives; we report these diagnostic histories in Appendix A.13.

Table 2: Six-point ablation: win rate of the evaluated seat against three copies of control seed 957223 at the same update (256 games per cell, Wilson 95% intervals), and the paired difference between the arm's own bargainer and a random bargainer.

Evaluated seat	Update	Win rate %	Own − random bargainer
control seed 534875 (bridge)	50	25.4 [20.4, 31.1]	—
control seed 534875 (bridge)	100	27.3 [22.2, 33.1]	—
full-state arm	50	8.6 [5.7, 12.7]	−1.6 [−6.2, 3.1]
no-bridge arm	50	10.5 [7.4, 14.9]	+1.6 [−3.1, 6.2]
no-bridge arm	100	11.7 [8.3, 16.2]	+2.3 [−2.7, 7.4]

Matched evaluations show a corresponding performance loss. Against three copies of a value bridge-trained control at the same update, an independently trained value bridge agent wins 25.4%–27.3% of games, whereas the completed full-state and no-value bridge ablations win 8.6%–11.7%. Replacing either ablation's trained bargainer with a random bargainer changes win rate by only -1.6 to 2.3 percentage points, with every paired confidence interval including zero; thus, these evaluations detect no advantage of the ablation bargainers over random bargaining. Together, the results provide causal evidence, conditional on the evaluated runs, that the allocation-sensitive value bridge is important for learning selective and useful bargaining. Because each alternative contains only one training seed and uses the shorter six-point environment, the experiment does not estimate across-seed variability or establish that the same effect size holds in the ten-point league setting.

The raw learned value differences are substantially larger than the immediate PBRS increments during league training, although this magnitude comparison is descriptive because shaping helps train the critic that produces those differences (Appendix A.12).

5.5 Interpretation and Limitations

These results demonstrate that economically recognizable bargaining behavior can emerge through task-oriented self-play without negotiation demonstrations, prescribed exchange prices, or hand-coded acceptance thresholds. The claim is not that the agents recover equilibrium bargaining strategies or human preferences. Rather, the learned bargainers organize proposals, responses, and counterpart selection in ways that improve independently measured task performance.

Several limitations bound this conclusion. The principal results come from two league lineages, and each alternative ablation contains only one training seed in a shorter six-point environment. The paired intervals characterize gameplay uncertainty conditional on trained policies, not variability across training runs. Behavioral statistics are primarily descriptive, the random-bargainer baseline trades destructively rather than representing an absence of trade, and the evaluated opponents come from the trained populations rather than from established external agents or humans. The policy swaps also reveal no detectable improvement in the macro policy. A scripted-placement analysis identifies initial placement as an important remaining weakness, but does not establish absolute playing strength or distance from optimal play; full placement, duration, reward-scale, and historical training diagnostics appear in Appendices A.10, A.11, A.12, and A.13.

6 Conclusion

Two independently trained league lineages show that agents can acquire useful and economically recognizable bargaining behavior through hierarchical self-play. Later hierarchies improve against common opponents, and paired policy swaps attribute the evaluated gains to the bargaining policy rather than to detectable changes in the macro policy. As training proceeds, agents convert more proposals into agreements, rely more heavily on strategic counterparties, and make fewer but more successful counteroffers. In a separate six-point experiment, replacing the allocation-sensitive value bridge with a full-state value difference or completion incentive alone produces near-universal agreement and substantially lower matched performance. Although the evidence is limited to two principal lineages and single-seed ablation alternatives, the results support the central framework: learn how resource allocations affect continuation in the surrounding task, and use that learned signal to train bargaining without demonstrations or prescribed exchange values.

AI USE STATEMENT

In this work, we used generative AI tools to help implement methods: an AI coding assistant helped develop the initial Python implementation of the methods explored here and provided a first draft of the Settlers of Catan environment implementation; all generated code was reviewed, modified, and tested by the authors. Additionally, we used generative AI tools to write batch-job shell scripts for running experiments, to write analysis and plotting code over existing experiment records, to create and modify figures, to identify relevant literature and format references, and to edit parts of the paper for readability and LaTeX formatting. The theoretical framework, hypotheses investigated, experimental methodology, and interpretation of results were done exclusively by the authors.

We have reviewed all AI-assisted work: reported numbers were checked against the original experiment records, suggested references were verified against their published sources, and AI-suggested text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83548–83599, 2024. doi: 10.52202/079017-2658. URL https://arxiv.org/abs/2309.17234.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1806.07857.

Ken Binmore, Ariel Rubinstein, and Asher Wolinsky. The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, 17(2):176–188, 1986. doi: 10.2307/2555382. URL https://doi.org/10.2307/2555382.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z. Leibo, Karl Tuyls, and Stephen Clark. Emergent communication through negotiation. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.03980.

Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):9737–9757, 2025. ISSN 2162-2388. doi: 10.1109/tnnls.2024.3497992. URL http://dx.doi.org/10.1109/TNNLS.2024.3497992.

Heriberto Cuayáhuitl, Simon Keizer, and Oliver Lemon. Strategic dialogue management via deep reinforcement learning. NIPS Deep Reinforcement Learning Workshop, 2015. URL https://arxiv.org/abs/1511.08099.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018. URL https://arxiv.org/abs/1709.04326.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback, 2023. URL https://arxiv.org/abs/2305.10142.

Quentin Gendre and Tomoyuki Kaneko. Playing catan with cross-dimensional neural network. In Haiqin Yang, Kitsuchart Pasupa, Andrew Chi-Sing Leung, James T. Kwok, Jonathan H. Chan, and Irwin King (eds.), *Neural Information Processing: 27th International Conference, ICONIP 2020*, volume 12533 of *Lecture Notes in Computer Science*, pp. 580–592, Cham, 2020. Springer. doi: 10.1007/978-3-030-63833-7_49. URL https://doi.org/10.1007/978-3-030-63833-7_49.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1256. URL https://aclanthology.org/D18-1256/.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.05397.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, and Oliver Lemon. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 480–484, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2077/.

Vijay Krishna and Roberto Serrano. Multilateral bargaining. *The Review of Economic Studies*, 63 (1):61–80, 1996. doi: 10.2307/2298115. URL https://doi.org/10.2307/2298115.

Tejas D. Kulkarni, Karthik R. Narasimhan, Ardavan Saeedi, and Joshua B. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1604.06057.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259/.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287. Morgan Kaufmann, 1999.

Martin J. Osborne and Ariel Rubinstein. *Bargaining and Markets*. Academic Press, San Diego, CA, 1990. ISBN 9780125286329.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982. doi: 10.2307/1912531. URL https://doi.org/10.2307/1912531.

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, 2017. URL https://arxiv.org/abs/1707.06347.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999. doi: 10.1016/S0004-3702(99)00052-1. URL https://doi.org/10.1016/S0004-3702(99)00052-1.

István Szita, Guillaume Chaslot, and Pieter H. M. Spronck. Monte-carlo tree search in settlers of catan. In H. Jaap van den Herik and Pieter Spronck (eds.), *Advances in Computer Games*, volume 6048 of *Lecture Notes in Computer Science*, pp. 21–32. Springer, Berlin, Heidelberg, 2010. doi: 10.1007/978-3-642-12993-3_3. URL https://doi.org/10.1007/978-3-642-12993-3_3.

Robert Shaun Thomas. *Real-time Decision Making for Adversarial Environments Using a Plan-based Heuristic*. Ph.d. thesis, Northwestern University, 2003.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3540–3549. PMLR, 2017. URL https://proceedings.mlr.press/v70/vezhnevets17a.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://doi.org/10.1038/s41586-019-1724-z.

Konstantia Xenou, Georgios Chalkiadakis, and Stergos D. Afantenos. Deep reinforcement learning in strategic board game environments. In Marija Slavkovik (ed.), *Multi-Agent Systems: 16th European Conference, EUMAS 2018, Bergen, Norway, December 6–7, 2018, Revised Selected Papers*, volume 11450 of *Lecture Notes in Computer Science*, pp. 233–248. Springer, Cham, 2019. doi: 10.1007/978-3-030-14174-5_16. URL https://doi.org/10.1007/978-3-030-14174-5_16.

A Appendix

Reproducibility statement

The accompanying exports contain the two league lineages' 8,640 tournament episode records and 144 matchup counts, their 900 reward-scale summaries (bounded at update 450), the per-episode and per-seat records of the 5,120 policy-swap games, the 512 mixed-bargainer games, and the 4,096 scripted-placement games, and the six-point ablation evaluation records. The historical diagnostic is separately traceable to the original repository's log file. Source hashes and plotting code document provenance, and seed manifests specify shared evaluation randomization. Original training manifests record modified working trees, so exact retraining additionally requires the corresponding runtime patches.

A.1 Additional Architecture Details

Agents are trained with Actor-Critic style PPO (Schulman et al., 2017), using generalized advantage estimation (Schulman et al., 2016). The macro policy $\pi_{M,i}(\cdot)$, macro critic $V^{\pi}_{M,i}(\cdot)$, and allocation-sensitive auxiliary state value function used in the value bridge $C_{\alpha,i;\theta_\alpha}(\cdot)$ are modeled as neural networks with disjoint parameter sets; the bargainer policy $\pi_{B,i}(\cdot)$ and critic share an encoder backbone.

The full macro state observation in our Catan experiments combines a 453-dimensional board representation, four 18-dimensional player summaries, an 11-dimensional context vector, Bayesian belief-state probabilities for private card ownership, and recent-trade history. Separate policy and macro critic encoders combine modality-specific networks and a trade-history GRU into 512-dimensional representations.

The allocation-sensitive auxiliary state value function used in the value bridge uses a separate, layer-normalized network with a 31-dimensional input. This input includes 13 ego/context dimensions and six dimensions for each opponent agent. The ego features describe asset holdings, private card count, public victory points, construction, and turn progress; each opponent contributes five resource counts and their public victory points. This representation concentrates the critic's input on quantities relevant to resource exchange, while omitting board geometry.

The bargaining policy combines participant and context MLPs, four-head attention over opponents and pending offers, and a GRU over eight recent offers. A 256-dimensional fused representation feeds a scalar bargaining critic and 13 categorical action heads: a decision head, a target head, ten resource-amount heads, and a pending-offer index head. Sequential feasibility masks condition the sampled components on earlier choices.

A.2 Auxiliary Return and Regression Loss

The return associated with the rewards provided to the auxiliary state value function used in the value bridge is

$$G_{\alpha,i,t} = \sum_{\ell=0}^{T-t-1} \gamma_M^{\ell} r^{(\alpha)}_{M,i}(s_{t+\ell}, a_{t+\ell}, s_{t+\ell+1}).$$

We model the auxiliary state value as a differentiable function with parameters θ_α, and indicate this dependence by writing $C_{\alpha,i;\theta_\alpha}(\psi_i(s_t))$. Its loss on a minibatch of transitions $\mathcal{D}$ is

$$\mathcal{L}_\alpha(\theta_\alpha) = \frac{1}{2|\mathcal{D}|} \sum_{(i,t)\in\mathcal{D}} \left[C_{\alpha,i;\theta_\alpha}(\psi_i(s_t)) - G_{\alpha,i,t}\right]^2.$$

In practice, we construct bootstrapped targets with generalized advantage estimation (Schulman et al., 2016).

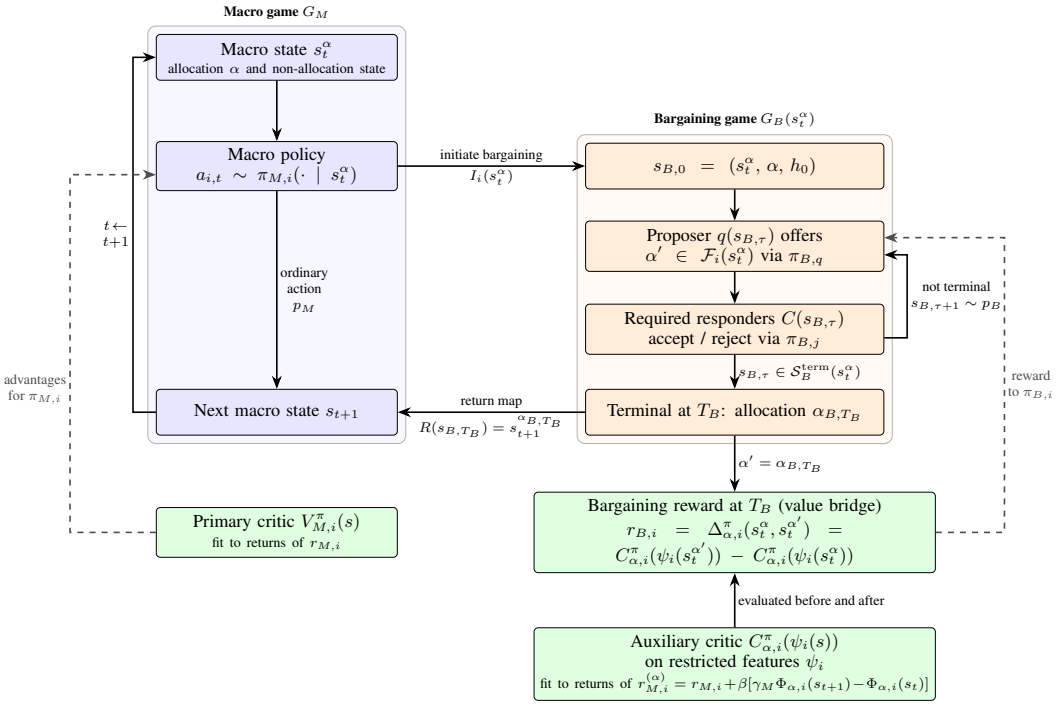

Figure 2: Linked macro and bargaining games with the allocation-sensitive value bridge. The macro policy either takes an ordinary action, with transitions governed by p_M, or initiates bargaining through the initiation map I_i. In the bargaining game, the active proposer offers a feasible reallocation, the required counterparties accept or reject, and play continues under p_B until a terminal state at T_B; the return map R resumes the macro game at $s_{t+1}^{\alpha_{B,T_B}}$. The primary critic $V_{M,i}^\pi$ is fit to macro-game returns and supplies the macro policy's advantages. The auxiliary critic $C_{\alpha,i}^\pi$ is fit to the shaped return on the restricted features $\psi_i(s)$, and its difference between the allocations before and after bargaining is the bargainers' terminal reward $r_{B,i}$. Solid arrows are state flow; dashed arrows are learning signals.

Table 3: Optimization settings, identical for the league lineages and the six-point ablation runs.

Setting	Macro	Bargainer
Discount	0.995	0.99
GAE parameter	0.97	0.95
PPO clip range	0.2	0.2
Epochs per update	3	2 maximum
Minibatch size	512	512
Entropy coefficient	0.10	0.04
Value-loss coefficient	0.5	0.5
Actor learning rate	10^{-4}	3×10^{-6}
Primary-critic learning rate	10^{-4}	Shared optimizer
Resource-critic learning rate	10^{-3}	—
Gradient-norm clip	1.0 per branch	0.5
Approximate KL stopping threshold	—	0.3
AdamW weight decay	10^{-4}	10^{-4}
AdamW numerical epsilon	10^{-5}	10^{-5}
Mixed precision	FP16	FP16

A.3 HYPERPARAMETERS

A.4 REWARD IMPLEMENTATION

The game-reward component is based on victory-point progress. At a reward check, let v_i^- be the player's previously recorded total victory points, v_i^+ its current total, and $W = 10$ the winning threshold. The point-based contribution is

$$r_{M,i}^{\text{check}} = \begin{cases} 10, & v_i^- < W \le v_i^+, \\ \min(v_i^+, W) - \min(v_i^-, W), & \text{otherwise.} \end{cases} \tag{1}$$

Thus the threshold-crossing reward replaces, rather than supplements, the ordinary point increment. Transitions without a point-reward check contribute no additional point reward. The evaluated configuration has no negotiation-attempt cost or failed-negotiation penalty.

The shaping potential measures partial fulfillment of construction costs. Let $\mathcal{B}$ contain roads, settlements, cities, and development cards, let c_{bk} be the cost of resource k for build type b, and set $R_b = \{k : c_{bk} > 0\}$. For holdings α_i, define

$$\Phi_i(\alpha_i) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{1}{|R_b|} \sum_{k \in R_b} \sqrt{\min\left(\frac{\alpha_{i,k}}{c_{bk}}, 1\right)}. \tag{2}$$

In resource order (wood, brick, wheat, sheep, ore), the cost vectors are

$$\begin{aligned} c_{\text{road}} &= (1,1,0,0,0), \\ c_{\text{settlement}} &= (1,1,1,1,0), \\ c_{\text{city}} &= (0,0,2,0,3), \\ c_{\text{development}} &= (0,0,1,1,1). \end{aligned}$$

The potential lies in $[0,1]$ and measures resource readiness, not whether a legal building location or piece is available. Its definition and weight are held fixed throughout training. It encodes construction requirements without specifying player-exchange prices, proposal targets, or acceptance thresholds.

A.5 DATA COLLECTION

Collection uses 20 NVIDIA T4 ranks with 18 environments per rank and a 512-transition macro horizon. Each nominal update therefore collects 184,320 learner macro transitions. Rank zero optimizes the gathered trajectories and broadcasts parameters. A complete bargaining invocation occupies one stored macro transition, independently of its number of internal bargaining events; bargainer returns use their own decision clock. These transition counts exclude opponent actions and internal bargaining decisions.

A.6 TRAINING LEAGUE

The league maintains separate macro and bargaining snapshot pools. Opponent macro and bargainer policies are sampled separately, so an opponent's two policies need not come from the same update during training. Once the pools are populated, each policy slot has a 0.1 probability of using the current policy; otherwise it draws a frozen opponent using a variance-weighted version of prioritized fictitious self-play.

Base snapshot-pool recipe The base recipe adds snapshots every ten updates and retains up to 20 recent snapshots, alongside at most 60 historical anchors saved every 100 updates. An ordinary pool draw selects the anchor branch with probability 0.2 and samples uniformly within that branch. Recent snapshots use the legacy difficulty-weighted sampler, whose raw weight is proportional to $\exp(-m_j)$, where m_j is an exponentially weighted learner-outcome score. The base recipe therefore already uses prioritized opponents; the six-point control and ablation runs of Section 5.4 use it unchanged.

Variance-weighted PFSP The league extension uses prioritized fictitious self-play, following the population-training motivation of AlphaStar (Vinyals et al., 2019). For a snapshot tag j, an outcome report $z \in \{-1, 0, 1\}$ denotes a learner loss, draw, or win. The rank-local score and raw sampling weight are

$$m_j \leftarrow 0.95 m_j + 0.05 z, \qquad m_j^{(0)} = 0,$$
$$\widehat{p}_j = \mathrm{clip}\left(\frac{m_j + 1}{2}, 0, 1\right), \qquad w_j = \widehat{p}_j(1 - \widehat{p}_j). \tag{3}$$

Here $\widehat{p}_j$ is a draw-adjusted learner-outcome estimate, not a pairwise win probability against j alone: the same four-player game outcome is reported for each pool-controlled opponent slot. Completed training games update these scores on the collecting rank; scores are not globally averaged across ranks. Routine pool-based evaluation also reports outcomes to its local pool.

Within the recent-plus-exploiter candidate set $\mathcal{P}$, the implemented distribution is

$$\widetilde{q}_j = \max\left(\frac{w_j}{\sum_{k \in \mathcal{P}} w_k}, 0.05\right), \qquad q_j = \frac{\widetilde{q}_j}{\sum_{k \in \mathcal{P}} \widetilde{q}_k}. \tag{4}$$

All-zero weights fall back to uniform sampling. The 0.05 floor is applied *before* the final renormalization, so it is not a guaranteed minimum final probability. This weighting emphasizes intermediate observed outcomes rather than exclusively sampling the hardest opponents.

Anchors, reference draws, and admission League anchors are saved every 50 updates, with a capacity of 60. Conditional on reaching the ordinary pool sampler, the anchor branch has probability 0.4 and samples uniformly; the other branch uses Eq. (4) over up to 20 recent snapshots and six exploiter snapshots. The macro pool also performs two sequential reference-draw tests before this branch choice: probability 0.05 for a uniform-legal random macro policy, then, if that test fails, probability 0.05 for the oldest available anchor. The bargaining pool does not perform these two tests. These are training opponents, not an unseen test set, despite the configuration's "holdout" naming. Fallbacks handle initially empty pools and unavailable anchors.

Regular snapshots are considered every ten updates. A candidate is admitted if it wins at least 25% of 36 stochastic games against three copies of the most recently admitted regular checkpoint hierarchy. The first snapshot and every anchor tick bypass this gate. Admission accepts or rejects the macro and bargaining snapshots together; rejection does not roll back the live learner. This gate is a pool-management heuristic, not a statistical guarantee of improvement. Its games are distinct from the reported checkpoint tournaments.

Exploiter opponents Each league lineage has a separate exploiter learner: seed 184674 targets main seed 221099, and seed 993645 targets main seed 408751. Exploiters train against frozen copies of their target main hierarchy, checking for a newer target checkpoint every ten exploiter updates. Every 25 exploiter updates they donate macro and bargaining snapshots to the target pools; donation logs confirm this mechanism operated. The main pools periodically rescan for these external snapshots and retain at most six exploiter tags per pool. These two auxiliary learners are not counted as additional independently evaluated main-agent lineages. Their training interactions and donation/admission evaluations are not included in the tournament game totals below.

A.7 Experimental Design

We evaluate learned policies in fixed tournaments with no parameter updates. Whenever two policies or experimental conditions are compared, we use common random numbers so that corresponding games share board generation, player order, dice rolls, and other exogenous events. We report paired bootstrap confidence intervals over games. Full tournament schedules, random seeds, bootstrap procedures, and treatment of draws are provided in Appendix A.8.

Checkpoint tournaments For each league lineage, we evaluate the complete macro–bargaining hierarchy every 50 training updates from update 50 through update 450. Each checkpoint matchup contains 60 games, and both components of every opponent are drawn from the same checkpoint. To measure within-lineage progress, we compare the update-50 and update-450 agents against the

same set of intermediate opponents, yielding 420 paired games per candidate. These comparisons measure progress against a common opponent population rather than uncertainty across additional training seeds.

Paired policy and placement experiments The policy-swap, mixed-bargainer, and scripted-placement experiments compare policies against three fixed opponents over 256 paired games per condition. Common randomization holds the exogenous random inputs fixed across conditions and balances the evaluated agent across the four starting positions. These experiments isolate the effects of changing the macro policy, bargaining policy, or initial-placement policy while leaving the evaluation environment and opponents unchanged.

Value-bridge ablation We conduct a separate six-victory-point experiment comparing three bargaining reward designs: the proposed allocation-sensitive value bridge, the primary macro critic's pre/post-bargaining value difference, and no learned value difference beyond a completion bonus. Three independently trained value-bridge runs serve as controls, while each ablation contains one fresh training seed. Policies are evaluated in paired 256-game comparisons against fixed control opponents; random-bargainer and independently trained bridge policies provide additional reference conditions. Because the ablation alternatives contain only one training seed each, these comparisons characterize the resulting trained policies but do not estimate variability across training runs.

A.8 Evaluation and Statistical Procedures

Common Randomization and Uncertainty Estimates All policy evaluations are conducted in fixed tournaments with learning disabled. Within each paired comparison, games use common random numbers: the evaluation root seed and episode index jointly determine separate random streams for board generation, player order, dice rolls, development cards, and other stochastic events. Corresponding episodes therefore share the same exogenous realization across experimental conditions, although their endogenous trajectories may diverge as the evaluated policies take different actions.

Unless otherwise noted, draws are coded as non-wins when calculating win rates. Confidence intervals are computed by resampling paired episode indices rather than treating observations from different conditions as independent. The reported intervals are pointwise, are not adjusted for multiple comparisons, and are conditional on the trained policies and selected opponent populations. They quantify evaluation uncertainty arising from stochastic gameplay but not variability across additional training seeds.

Checkpoint Tournaments We evaluate each league lineage at updates $50, 100, \ldots, 450$. Both the macro and bargaining policies of an evaluated agent are taken from the same checkpoint. Opponents are likewise evaluated as complete checkpoint hierarchies: their macro and bargaining policies are not independently crossed during these tournaments.

Each populated checkpoint matchup contains 60 games. For each lineage, the resulting tournament matrix contains 72 populated matchups and therefore 4,320 games. Across the two league lineages, the checkpoint tournaments contain 8,640 games, of which 8,625 produce a winner and 15 end in draws. All checkpoint tournaments use evaluation root seed (777001).

To compare the update-50 and update-450 agents within a lineage, we restrict both candidates to the common opponent checkpoints from updates 100 through 400. The seven common opponents yield $7 \times 60 = 420$ games for each candidate. Because the two candidates face the same opponent checkpoint under the same 60 episode indices, the comparison is paired by both opponent and exogenous game realization.

We construct 95% confidence intervals for the early-to-late win-rate difference using 20,000 paired cluster-bootstrap replicates with analysis seed (20260907). Each replicate resamples the 60 episode indices and applies the same sampled indices jointly across the seven common opponents. Clustering by episode index preserves the common-random-number dependence across opponent matchups.

Policy-Swap, Mixed-Bargainer, and Placement Experiments The policy-swap, mixed-bargainer, and scripted-placement experiments use a common paired design. In each experiment, the evaluated agent faces three fixed opponents. Every experimental cell contains 256 games gen-

Table 4: Complete hierarchies against common opponents (updates 100–400), both league lineages. Rates are percentages; gains and paired 95% intervals are percentage points; 420 games per candidate.

Seed	Updates	Early	Late	Gain [95% interval]
221099	50 / 450	16.4	28.8	$+12.4$ $[5.5, 19.5]$
408751	50 / 450	14.5	23.6	$+9.0$ $[2.6, 15.5]$

erated from evaluation root seed (424242). The evaluated agent occupies each of the four starting positions in exactly 64 games.

For a given episode index, all experimental cells receive the same board, player order, dice sequence, development-card deck, and other exogenous random variables. Comparisons between cells therefore estimate within-episode effects of changing the macro policy, bargaining policy, or placement policy. We construct 95% confidence intervals from 10,000 paired bootstrap replicates over the 256 episode indices. Draws are coded as non-wins; five of the 5,120 policy-swap games end in draws.

Six-Point Value-Bridge Ablation The value-bridge ablation uses a six-victory-point version of the environment to reduce the computational cost of training multiple conditions. The experiment compares three bargaining reward designs:

1. **Allocation-sensitive bridge.** The bargaining policy receives the normalized and clipped pre/post-bargaining difference from the allocation-sensitive auxiliary critic.
2. **Full-state difference.** The allocation-sensitive bridge is replaced by the corresponding pre/post-bargaining difference from the primary full-state macro critic.
3. **No learned bridge.** The learned value-difference reward is removed, leaving only the bargaining-completion bonus.

Three completed allocation-sensitive-bridge runs, with training seeds (957223), (534875), and (169504), serve as controls. The full-state-difference and no-bridge conditions each contain one independently trained run, with seeds (156449) and (509782), respectively. Except for the bargaining reward source, the ablation runs use the same six-point environment, architecture, optimization settings, and training procedure as the corresponding controls.

At each evaluated training update, an ablation policy plays 256 paired games against three copies of the allocation-sensitive control trained with seed (957223). Evaluations use root seed (515151). We additionally evaluate each trained macro policy with a random bargaining policy to measure dependence on its learned bargainer. The independently trained allocation-sensitive control with seed (534875) provides a second bridge-trained reference, allowing the observed differences between ablation arms and the primary control to be compared with variation between two bridge-trained policies.

All ablation comparisons use paired episode indices and common exogenous randomization. Because the two ablation conditions contain only one training seed each, their evaluation intervals measure stochastic game-level uncertainty conditional on the trained policies. They do not estimate the distribution of outcomes that would result from retraining each condition with new random seeds and should not be interpreted as across-seed significance tests.

A.9 Additional Results

A.10 Initial Placement: The macro policy's Remaining Weakness

A settlement's expected production is summarized by its *pips*, the number of dice combinations that roll each adjacent number token ($6 - |7 - n|$ for token n; 6 and 8 carry five pips, 2 and 12 one). We score every initial placement in the 5,120 swap games and regenerate each board from its seed to compute two references: a random legal snake draft and a greedy draft that always takes the max-pip legal node.

Table 5: Policy-swap win rates of the evaluated seat (percent, 256 games per cell). Panel A: own-lineage homogeneous 3xU400. Panel B: cross-lineage mixed table U150/U300/U400 (other seed).

Panel	Lineage	C1	C2	C3	C4	C5
A	221099	15.6	24.6	11.3	23.0	2.0
A	408751	14.8	21.1	14.1	22.3	2.7
B	221099	17.2	25.8	16.4	29.3	5.9
B	408751	18.8	28.9	13.3	28.1	2.3

Table 6: Scripted pip-greedy placement versus learned placement, paired win-rate effects of the evaluated update-400 hierarchy in percentage points [95% paired interval], 256 pairs per cell. Pips are the evaluated seat's two-settlement total.

Scripted seats, picks	Pips	221099 A	221099 B	408751 A	408751 B
learned (reference C4)	16.7	23.0%	29.3%	22.3%	28.1%
seat 0, both picks	24.6	$+26.6$ $[19.1, 34.0]$	$+18.0$ $[10.5, 25.8]$	$+34.8$ $[27.0, 42.2]$	$+26.2$ $[18.4, 34.0]$
seat 0, first pick	21.0	$+14.5$ $[7.4, 21.5]$	$+8.6$ $[1.6, 15.6]$	$+19.1$ $[11.7, 26.6]$	$+13.7$ $[5.9, 21.5]$
all four, both picks	21.0	$+3.5$ $[-3.9, 10.9]$	-4.7 $[-12.5, 3.1]$	-2.3 $[-9.8, 4.7]$	$+7.0$ $[-1.2, 15.2]$
all four, first pick	18.6	-0.4 $[-7.4, 6.6]$	-0.4 $[-8.2, 7.4]$	-0.8 $[-7.8, 6.2]$	$+3.9$ $[-4.3, 12.1]$

Learned placement is far from greedy The two initial settlements total 12.8 pips under the random draft, 15.8 for update-50 macro policies, 16.7 to 16.8 for updates 150 to 400, and 21.1 under the greedy draft. The update-400 macro policy takes the best available node on 9% of its picks, its chosen node ranks eighth by median among the 40 to 50 legal nodes, and as first mover it picks an 8.4-pip node when a 13.1-pip node is available. The paired update-50 to update-400 improvement is $+0.8$ to $+1.1$ pips in every cell, complete by update 150: the only macro policy improvement the data detect. The macro policy did learn ports: 81% of players open on a port and 47% on a 2:1 port, against 56% for random placement, although port access separates winners only weakly (85% of winners versus 80% of non-winners). Placement matters a great deal for the outcome: ranking the four seats within each game by opening pips, the highest seat wins 44%, the second 27%, the third 16%, and the lowest 8%; even at all-update-400 tables, seats opening with 22 or more pips win 54% and seats with 12 or fewer win 6%. Part of this gap is a design choice of ours in the environment interface. The macro observation encodes each tile's number token as a raw integer 0–12, which is non-monotonic in production: the two-dice total has a triangular distribution peaking at 7, so 6 and 8 carry five combinations each and 2 and 12 one. The macro policy therefore has to infer the dice distribution from sparse game reward rather than read production directly, and the evidence is that it did so only partially, whereas ports, which the observation marks explicitly, it learned to seek.

Scripted-placement interventions To measure what learned placement costs, we replace initial placement with the greedy rule while the update-400 macro policy and bargainer play from the first main turn, in four variants: the evaluated seat only or all four seats, and both setup picks or only the first pick (the second pick then returns to the learned macro policy). Every variant plays the 256 shared boards against Panels A and B and is paired with the C4 games. Table 6 reports the effects.

Scripting both of the evaluated seat's picks raises its win rate from 22–29% to 47–57%, a gain of 18 to 35 points, with every scripted pick verified to be the max-pip legal node. Scripting only its opening settlement recovers 9 to 19 points; the learned second pick then reaches 21.0 pips, still 3.6 short of the greedy 24.6. The 24.6 exceeds the 21.1 of an all-greedy draft because the learned opponents leave high-production intersections unclaimed, so the one-seat gains are a ceiling. When all four seats place greedily, every seat opens at 21.0 pips and all effects include zero, as a symmetric intervention should; the same holds for greedy first picks at all seats (18.6 pips). The one-seat first-pick cell is the cleanest statement: replacing only the opening settlement with the max-pip legal node raises the late agent's win rate by 9 to 19 points on identical boards, at the 21-pip level a fair greedy table also reaches. Greedy pips is a production ceiling, not an optimal placement rule. Good openings sometimes give up pips for resource diversity, a 2:1 port, the ore–wheat pairing that cities require, or an expansion route, and the fully greedy table does not beat the learned one. The intervention therefore measures the cost of under-weighting production at the opening, not the distance from the learned macro policy to optimal placement.

Table 7: Magnitude diagnostics over the first 450 training updates of each league lineage. RMS columns are medians over updates; ratios are medians of per-update ratios. "Last 50" denotes the last 50 analyzed updates.

Seed	Updates	Delta RMS	PBRS RMS	Ratio	Last 50
221099	450	0.149	0.0268	5.53	5.32
408751	450	0.148	0.0267	5.52	5.34

A.11 Game Duration

We measure duration by the environment's dice-roll counter. Across the two league tournaments, the 8,625 games with a winner required a median of 99 rolls (interquartile range 82–123, mean 112.2); restricted to matchups between updates 400 and 450, 239 games had a median of 89 rolls (76.5–111, mean 97.9). In the swap experiment the update-400 mirror tables (C4 on Panel A) have a median of 92 rolls and means of 102 to 108, and over all 5,115 decided swap games the median is 96 with 10th and 90th percentiles of 66 and 155. Five of 5,120 games reached the step cap.

Duration depends on all four players, so we read it as an absolute measure of a table's efficiency rather than of one agent's strength, and the placement interventions show how much of it is set at the opening. When the evaluated seat places greedily on both picks, its wins take 82 to 88 rolls against 107 to 111 for the learned agent's own wins, which are the slowest games in the data: the learned agent wins by outlasting the table rather than by racing to ten points. When all four seats place greedily, the whole table finishes about 10% sooner, 90 to 96 rolls on average (medians 82–88) against 102 to 108 (median 92), with wins and losses equally long. The shortest games in the study are these fully greedy tables: the own-lineage table of seed 408751 has a median of 82 rolls, against 92 on the same boards under learned placement, and the one-seat greedy cells of that lineage have medians of 83; greedy first picks alone barely change pace, so the second settlement is what speeds up the economy. Conversely, tables of early bargainers play 143 to 148 rolls. Bargaining quality and placement quality therefore both act on pace, in opposite directions from the learned agent's current setting. We do not compare these durations with human play, for which we have no matched measurement.

A.12 Scale of Learned Value Feedback Relative to PBRS

We examine the first 450 training updates of each league lineage, matching the evaluation cutoff. For each update and each signal, RMS magnitude is reconstructed as the square root of the squared logged mean plus the squared population standard deviation, and we divide the RMS of raw session value differences by the RMS of actual weighted immediate PBRS increments, before bargaining-reward normalization and clipping. The median per-update ratios are 5.53 and 5.52 (Table 7); the raw value difference exceeds the PBRS increment in 898 of 900 updates, the exceptions being the first update of each lineage, and the last-50-update medians are 5.32 and 5.34. Retaining PBRS therefore does not imply that its immediate increment is larger than the learned feedback. PBRS helps train the very critic producing those differences, so this diagnostic does not establish causal independence from shaping; the ablation of Section 5.4 addresses causality directly.

A.13 Bargaining failure early in development

Earlier development runs show the same loss of selectivity in the other direction. In the January 2026 repository, job 217169 resumed a dense-shaped bargainer at update 101 with a full-state critic difference (a second value head on the macro encoder, unnormalized, without shaping or completion bonus) as its only reward; the training agreement rate fell from 60.7% at update 102 to 3.6% at update 180, and its update-150 evaluation reported 0.9% success despite 479.5 attempts per game (Figure 3). A second full-state attempt on February 26, trained fresh with normalization and shaping but without a completion bonus, stayed at 4% agreement for 100 updates; its raw per-trade critic difference had a standard deviation of 0.004 to 0.007 against 0.12 to 0.16 for the resource critic in the runs that followed. Without an allocation-sensitive value signal the acceptance decision has nothing to optimize, and the bargainer drifts to whichever extreme the residual reward favors: never

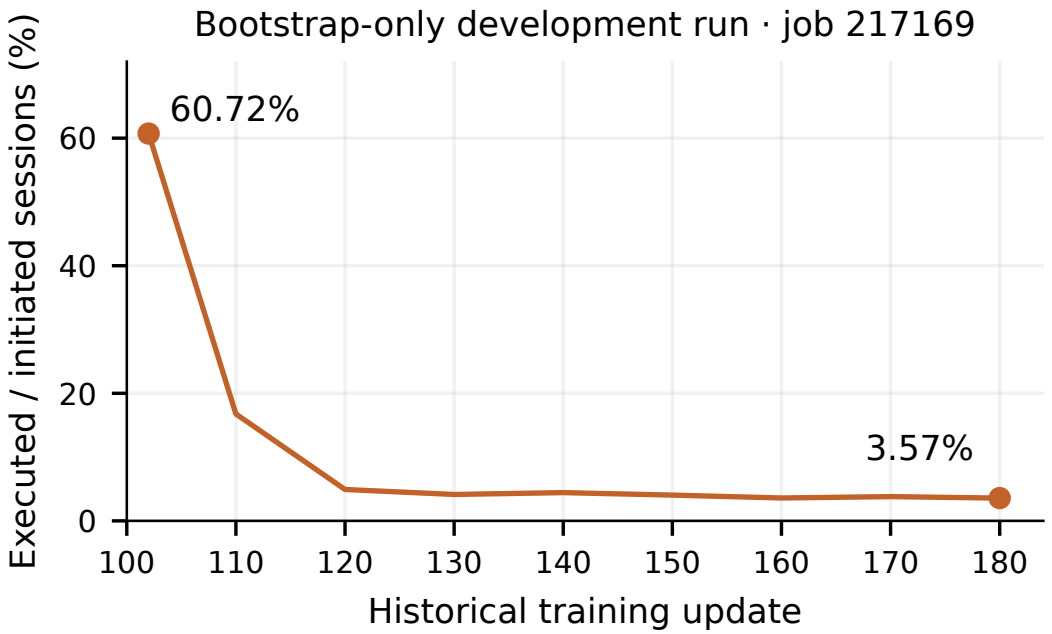

Figure 3: Development diagnostic from a single January 2026 job that rewarded the bargainer with a full-state critic difference only: unsmoothed logged training agreement rates collapse to near zero. It is an earlier six-point setting with different training settings and is not part of the matched ablation.

agreeing when that reward is zero, always agreeing when it is a small positive bonus. The controls carry the same bonus and remain selective.

B Future Work

The ablation should be extended to a second seed per arm, to the ten-point league setting, and to a shaping-free arm that separates the learned critic from its fixed potential; a matched comparison isolating PFSP, anchors, gating, and exploiter donations remains open. The placement result points to a concrete change: an explicit pip feature per tile and node in the macro observation, and a curriculum or auxiliary target for initial placement, followed by a repeat of the swap experiment to test whether the macro policy then contributes. Fixed heuristic opponents, greedy placement with a scripted build order and rule-based trade acceptance, would provide the external yardstick this evaluation lacks, and a hybrid of scripted placement with the learned policies would measure how much of the placement gap survives against opponents that also place well.

A second direction is behavioral generalization. Per-offer analyses stratified by game phase, resource scarcity, and outside-option rates could measure offer selectivity and exchange composition directly, replacing card counts with a valuation. Human comparisons would be needed to move from economically recognizable behavior to quantitative claims about similarity to human bargaining or human pace of play. Finally, the interface can be tested beyond Catan: a transferable-resource task with different downstream dynamics would assess whether learning a resource-sensitive continuation critic remains an effective way to train negotiation.