Learning in Causal Markov Games

Aurghya Maiti
Department of Computer Science
Columbia University
New York, NY 10025
aurghya@cs.columbia.edu

Elias Bareinboim
Department of Computer Science
Columbia University
New York, NY 10025
eb@cs.columbia.edu

Abstract

Markov Games are the standard formal model for multi-agent reinforcement learning, capturing agents that act in a shared state and optimize rewards over time. However, in real-world settings, agents' decisions are often influenced by unobserved factors, such as cognitive biases, behavioral tendencies, or intuitive signals, that also affect rewards and future states. Ignoring these unobserved confounders can make standard learning methods converge to suboptimal policies. In such settings, optimal play may require policies that condition on counterfactual signals, requiring reasoning at the counterfactual layer of the Pearl Causal Hierarchy. In this paper, we introduce Causal Markov Games (CMGs), a framework for modeling sequential multi-agent decision making in the presence of unobserved confounding. We show that CMGs strictly generalize Markov Games, with arbitrarily large gaps between classical interventional equilibria and causal counterparts. We then develop two learning algorithms under different observability assumptions. The first, CNash-VI-FO, is a model-based learner with finite-sample guarantees when agents' natural actions or intuitions are revealed post hoc. The second, CNash-VI-NO, is an explore-then-exploit procedure with asymptotic guarantees for the setting in which opponents' natural actions are never observed. To address scalability, we further provide a drop-in counterfactual augmentation of deep MARL. We also show empirically that counterfactual agents outperform their non-causal counterparts in the Iterated Causal Prisoner's Dilemma and earn higher mean reward on a confounded windy variant of a Multi-Particle Environment.

1 Introduction

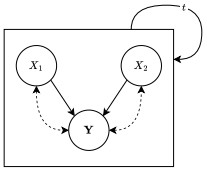
Sequential decision-making in multi-agent systems underlies a wide range of applications, from autonomous robots and traffic signal control to distributed resource management, large-scale game playing, and LLM-agent interaction. The standard model is the Markov Game (MG) [Shapley, 1953, Littman, 1994], where agents choose actions in a shared state that evolves stochastically and emits rewards. A substantial line of work studies *learning from samples* in this model: how agents converge to an equilibrium from finite interactions. Early work on Nash-Q [Hu and Wellman, 2003] and Friend-or-Foe Q-learning [Littman, 2001] established convergence to equilibrium under structural conditions. Later methods give finite-sample guarantees for Nash, correlated, and coarse-correlated equilibria in the game classes they treat: Nash value iteration and its optimistic variants [Liu et al., 2021], V-learning [Jin et al., 2021], and decentralized policy gradient [Daskalakis et al., 2020]. Xie et al. [2020] extend this line to function approximation.

However, in many practical settings, agents act on intuitions or private context that the learner cannot observe but that simultaneously influence the shared reward and next state. A physician's clinical judgment or an investor's market intuition acts as an *unobserved confounder* (UC): a latent factor that shapes both an agent's actions and the environment's response. Omitting such factors from the MG

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

		$X_2=0$		$X_2=1$	
		$U_2=0$	$U_2=1$	$U_2=0$	$U_2=1$
$X_1=0$	$U_1=0$	$-1.4,-1.4$	$-1,-1$	$-14,0$	$-8,0$
	$U_1=1$	$-1,-1$	$0,0$	$0,0$	$0,-3$
$X_1=1$	$U_1=0$	$0,-14$	$0,0$	$0,0$	$0,-8$
	$U_1=1$	$0,-8$	$-3,0$	$-8,0$	$0,0$

(a)



(b)

P_1	P_2	M_1	M_2
TfT	TfT	-1.0, -1.0	-1.0, -1.0
TfT	D	-1.9, -1.9	-1.9, -1.9
TfT	Int	-3.56, -0.08	-1.56, -4.28
D	Int	-2.4, 0	0, -8.9

(c)

Figure 1: (a) Y_1, Y_2 as functions of X_1, X_2, U_1, U_2; (b) causal diagram of the Iterated Causal Prisoner's Dilemma; (c) payoffs under different strategies; Int follows the natural intuition.

leaves the tacit knowledge underlying nearly all real multi-agent interaction outside what the MG and any learner built on it can express.

In the single-agent setting, the canonical answer is the *Markov Decision Process with Unobserved Confounders* (MDPUC) [Zhang and Bareinboim, 2022, Bareinboim et al., 2024]: an MDP augmented with exogenous latent variables that jointly drive transitions, rewards, and the behavior policy. MDPUC lifts decision-making to the counterfactual layer of the Pearl Causal Hierarchy (PCH) [Bareinboim, 2025, Bareinboim et al., 2022]; without this lift, even causal bandits incur unbounded regret under any observational or interventional algorithm [Bareinboim et al., 2015], with an analogous impossibility for MDPs. No such object exists in the sequential multi-agent case, and the gap is not notational, as the example below shows.

Example 1.1 (Iterated Causal Prisoner's Dilemma (ICPD))**.** Two friends repeatedly engage in criminal activity. After each offense they are apprehended and questioned separately. In each round $t \in \{1, \dots, H\}$, individual $i \in \{1, 2\}$ chooses $X_{i,t} \in \{0, 1\}$, either to remain silent (cooperate, $X_{i,t} = 0$) or to betray the other (defect, $X_{i,t} = 1$). Each round is shaped by latent circumstances $U_{i,t} \in \{0, 1\}$ that capture UCs such as the strength of evidence, the competence of legal counsel, and the disposition of the judge. These are unobserved by the agents but affect both their behavior and the outcome, with $P(U_{i,t} = 0) = 0.6$ independently across agents and rounds. Each individual also has an intrinsic ability $R_{i,t} \in \{0, 1\}$ for reading these circumstances, yielding the natural action $X_{i,t} \leftarrow f_X(R_{i,t}, U_{i,t}) = R_{i,t} \oplus U_{i,t}$ ($\oplus$ denotes exclusive-or); when $R_{i,t} = 1$ the agent cooperates iff circumstances are favorable; $R_{i,t} = 0$ reverses this judgment. The structural mechanisms and latent-variable distributions are fixed and unknown to the agents; latent realizations are drawn afresh each round. The per-round outcome $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})$ follows the payoff structure in Fig. 1a (causal graph in Fig. 1b), and agent i's overall utility is $(1/H) \sum_{t=1}^{H} Y_{i,t}$.

We compare two environments that are indistinguishable from the perspective of interventional actions. In environment M_1, both individuals always read their circumstances correctly ($R_{1,t} = R_{2,t} = 1$); in M_2, both always misread them ($R_{1,t} = R_{2,t} = 0$). The marginal distribution of $(U_{1,t}, U_{2,t})$ is identical across rounds in both environments. The induced interventional stage game is the same Prisoner's Dilemma in both cases. Always-defect (D) is an interventional equilibrium strategy; we also compare against tit-for-tat (TfT) as a reactive baseline [Shoham and Leyton-Brown, 2008].

The two environments diverge sharply once natural intuition is admitted. Mutual natural play yields per-round payoffs of $(0, 0)$ in M_1 but approximately $(-2.4, -2.4)$ in M_2; against tit-for-tat, a behavioral agent attains -0.08 while its opponent incurs -3.56 (simulated in M_1; Fig. 1c; derivations in Sec. 2 and Appendix F). In either environment, classical Markov-Game interventional equilibrium strategies are dominated by counterfactual play. □

These observations point to a missing object in the foundations of multi-agent reinforcement learning: a sequential game model that represents unobserved confounding, natural actions, interventions, and counterfactual policies within a single causal framework. Existing causal models address only parts of this problem (Sec. 1.2). We address this gap and introduce the *Causal Markov Game* (CMG), a causal generalization of Markov Games that lifts them to the PCH and makes it possible to define and learn counterfactual policies across stages. We make the following contributions:

1. We define the *Causal Markov Game* (CMG) and prove that it strictly generalizes the Markov Game (Thm. 2.4). Natural-policy payoffs can differ arbitrarily from interventional Nash

payoffs when reward magnitudes are unrestricted (Thm. 2.4); a bounded-reward example has a gap linear in the horizon between interventional and causal equilibria (Prop. C.3).

2. We develop two learning algorithms with provable guarantees: *CNash-VI-FO* (Sec. 3), a model-based algorithm that returns an ϵ-Causal Nash Equilibrium when natural actions are revealed post-hoc; without that observability, *CNash-VI-NO*, an explore-then-exploit variant for two-player general-sum CMGs whose exploitability converges to zero almost surely under certain identification assumptions (Sec. 3.2).

3. We provide a counterfactual augmentation for deep MARL (Sec. 3.3) and show empirically an advantage in the ICPD and higher mean reward for counterfactual agents on a Multi-Particle Environment.

1.1 Preliminaries

Notation. We use capital letters for random variables (X) and the corresponding lowercase letters for their values (x). Bold capitals ($\mathbf{X}$) and bold lowercase ($\mathbf{x}$) denote tuples of random variables and values, respectively. $\mathcal{D}_X$ denotes the domain of X, and $|\mathcal{S}|$ the cardinality of a set $\mathcal{S}$. We write $\Delta(\mathcal{A})$ for the set of probability distributions over a set $\mathcal{A}$.

Structural Causal Models. Our framework is built on Structural Causal Models (SCMs) [Pearl, 2009, Bareinboim, 2025]. An SCM is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{V}$ is the set of endogenous variables, $\mathbf{U}$ the set of exogenous (latent) variables, $\mathcal{F} = \{f_V\}_{V \in \mathbf{V}}$ a collection of structural assignments $V \leftarrow f_V(\mathrm{Pa}(V), \mathbf{U}_V)$ with $\mathrm{Pa}(V) \subseteq \mathbf{V}$ and $\mathbf{U}_V \subseteq \mathbf{U}$, and $P(\mathbf{U})$ a distribution over the exogenous variables. The model $\mathcal{M}$ induces the *observational* (or layer-L_1) distribution $P(\mathbf{V})$. For $\mathbf{X} \subseteq \mathbf{V}$, the intervention $\mathrm{do}(\mathbf{x})$ replaces $\{f_X : X \in \mathbf{X}\}$ with the constants $\mathbf{x}$, yielding the submodel $\mathcal{M}_{\mathbf{x}}$ and the *interventional* (layer-L_2) distribution $P_{\mathbf{x}}(\mathbf{Y}) = P(\mathbf{Y}_{\mathbf{x}})$. Given evidence $(\mathbf{x}', \mathbf{y}')$, the *counterfactual* (layer-L_3) distribution $P(\mathbf{Y}_{\mathbf{x}} \mid \mathbf{x}', \mathbf{y}')$ evaluates $\mathbf{Y}$ under $\mathrm{do}(\mathbf{x})$ in a world where $(\mathbf{x}', \mathbf{y}')$ was in fact observed. Together, L_1, L_2, and L_3 form the Pearl Causal Hierarchy (PCH); we refer the reader to [Bareinboim, 2025, Bareinboim et al., 2022] for the full formalism.

Actions with state information. We extend the action definitions of Bareinboim [2025], Maiti et al. [2025] from the single-step multi-agent setting to a stateful setting in which each agent observes a state $\mathbf{S}$ before acting. Let X_i denote the action variable of agent i, $\mathcal{A}_i$ its domain, and $\mathbf{U}_i$ the unobserved parents of X_i in $\mathcal{M}$. We define the two background action types here; the L_3 (counterfactual) action, which is the central object of this paper, is introduced in Sec. 2.

1. L_1 / Natural action: An L_1 action of agent i in state $\mathbf{s}$ is the value $X_i \leftarrow f_{X_i}(\mathbf{s}, \mathbf{U}_i)$ produced by the natural mechanism in $\mathcal{M}$. We denote the resulting *intuition* by $X_i' := f_{X_i}(\mathbf{s}, \mathbf{U}_i)$.

2. L_2 / Intervention: An L_2 action of agent i is a state-dependent hard intervention $\sigma_i : \mathcal{D}_{\mathbf{S}} \to \mathcal{A}_i$, executed as $\mathrm{do}(X_i = \sigma_i(\mathbf{s}))$ in state $\mathbf{s}$.

We write $\Pi_i^{(1)}$ and $\Pi_i^{(2)}$ for the corresponding policy spaces of agent i as distributions over L_1 and L_2 actions, respectively. Throughout the paper, $\mathcal{A}_i$ refers to an action *set* (the domain of X_i), and $\Pi_i^{(\cdot)}$ to a *policy class*; the counterfactual policy class Π_i^{CTF} is defined in Sec. 2.

1.2 Related Work

Work at the intersection of causal inference and sequential multi-agent learning is sparse. MD-PUC [Zhang and Bareinboim, 2022, Bareinboim et al., 2024] extends MDPs to include unobserved confounders and allow agents to perform counterfactual actions; however, their framework is limited to single-agent settings. The Causal Normal-Form Game (CNFG) [Maiti et al., 2025], on the other hand, incorporates counterfactual actions in a multi-agent system, but it only considers single-step games and does not consider learning with context. Partially observable stochastic games (POSGs) [Hansen et al., 2004] allow partially observed state. Here we focus on an observed Markov state with unobserved confounders and explicit natural-action and intervention mechanisms. Such confounding can also be present in POSGs, and we leave learning under such cases as future work.

2 Causal Markov Games

In this section, we introduce *Causal Markov Games*, a framework for sequential multi-agent decision-making in environments with unobserved confounders. We define the environment as an SCM (Def. 2.1) and introduce the *counterfactual policy* that lets each agent act on its own intuition (Def. 2.2); we illustrate each construction on the Iterated CPD of Ex. 1.1, and close with Thm. 2.4, which shows that the resulting framework strictly generalizes the Markov game.

Definition 2.1 (Causal Markov Game)**.** A *Causal Markov Game* (CMG) for n agents over horizon $H \in \mathbb{N}^+$ is a structural causal model

$$\mathcal{M} = \langle \mathbf{U},\, \mathbf{V},\, \mathcal{F},\, P(\mathbf{U}) \rangle, \tag{1}$$

whose components are as follows.

1. *Exogenous variables.* $\mathbf{U} = \{U_t\}_{t=1}^{H} \cup \{U_{s,H+1}\}$, where each $U_t = (U_t^C,\, \mathbf{U}_{a,t},\, U_{s,t},\, \mathbf{U}_{y,t})$ consists of a common latent variable U_t^C, agent-specific latent variables $\mathbf{U}_{a,t} = (U_{1,t}, \ldots, U_{n,t})$, a state latent $U_{s,t}$, and reward latents $\mathbf{U}_{y,t} = (U_{y_1,t}, \ldots, U_{y_n,t})$.

2. *Endogenous variables.* $\mathbf{V} = \mathbf{S} \cup \mathbf{X} \cup \mathbf{Y}$, with states $\mathbf{S} = (S_t)_{t=1}^{H+1}$, joint actions $\mathbf{X} = (\mathbf{X}_t)_{t=1}^{H}$ where $\mathbf{X}_t = (X_{1,t}, \ldots, X_{n,t})$, and rewards $\mathbf{Y} = (\mathbf{Y}_t)_{t=1}^{H}$ where $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{n,t})$. Here $X_{i,t}$ and $Y_{i,t}$ denote the action and reward of agent i at stage t.

3. *Structural assignments.* For each $t \in \{1, \ldots, H\}$ and $i \in \{1, \ldots, n\}$,

$$\begin{aligned} S_{t+1} &\leftarrow f_S(S_t,\, \mathbf{X}_t,\, \mathbf{U}_{a,t},\, U_t^C,\, U_{s,t+1}), \\ X_{i,t} &\leftarrow f_{X_i}(S_t,\, U_{i,t},\, U_t^C), \\ Y_{i,t} &\leftarrow f_{Y_i}(S_t,\, \mathbf{X}_t,\, \mathbf{U}_{a,t},\, U_t^C,\, U_{y_i,t}). \end{aligned} \tag{2}$$

4. *Exogenous distribution.* The U_t are i.i.d. across t, independent of $U_{s,H+1}$, with $P(U_t) = P(U_t^C)\, P(U_{s,t}) \prod_{i=1}^{n} P(U_{i,t})\, P(U_{y_i,t})$; hence the $2n+2$ components of each U_t are mutually independent.

The model is initialized with the state S_1. □

The structural form in Eq. (2) encodes a few design choices that distinguish a CMG from a Markov game. Agent i's natural mechanism f_{X_i} reads its own latent $U_{i,t}$ and the shared latent U_t^C, while the reward mechanism f_{Y_i} may read the full vector $(\mathbf{U}_{a,t}, U_t^C)$, so actions and rewards are confounded. The state S_t depends on the previous stage's $(\mathbf{U}_{a,t-1}, U_{t-1}^C)$, so actions and the next state are confounded as well, and U_t^C confounds the agents' actions with one another. Define the intuition as $X'_{i,t} := f_{X_i}(S_t, U_{i,t}, U_t^C)$. Two structural consequences follow. First, mutual independence of $\{U_{i,t}\}_{i \in [n]}$ yields the conditional independence of intuitions,

$$X'_{i,t} \;\perp\!\!\!\perp\; X'_{j,t} \;\big|\; (S_t, U_t^C) \qquad \text{for all } i \neq j, \tag{3}$$

so that, given the state, the intuitions form a correlated type space in the context of Bayesian games, with U_t^C as the common hidden component. Second, the latents $(\mathbf{U}_{a,t}, U_t^C)$ are used in both f_{X_i} and f_{Y_i}, so the observational and interventional reward distributions need not coincide, $P(\mathbf{y}_t \mid s_t, \mathbf{x}_t) \neq P_{\mathbf{x}_t}(\mathbf{y}_t \mid s_t)$, and an algorithm restricted to interventional reasoning cannot identify, let alone exploit, the dependence of behavior on $\{U_{i,t}\}$. A standard interventional policy ignores agent i's intuition entirely. We now define the policy class that lets each agent perceive its intuition and act on it.

Definition 2.2 (Counterfactual Policy)**.** For a CMG $\mathcal{M}$, the *counterfactual policy space* of agent i is

$$\Pi_i^{\mathrm{CTF}} \;=\; \big\{ \pi_i = (\pi_{i,1}, \ldots, \pi_{i,H}) \;:\; \pi_{i,t}(X_{i,t} \mid S_t, X'_{i,t}) \in \Delta(\mathcal{A}_i) \big\}, \tag{4}$$

where each decision $\pi_{i,t}$ maps the state S_t and the agent i's intuition $X'_{i,t}$ to a distribution over executed actions $X_{i,t} \in \mathcal{A}_i$. A *joint counterfactual policy* is a tuple $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \in \prod_{i=1}^{n} \Pi_i^{\mathrm{CTF}}$.

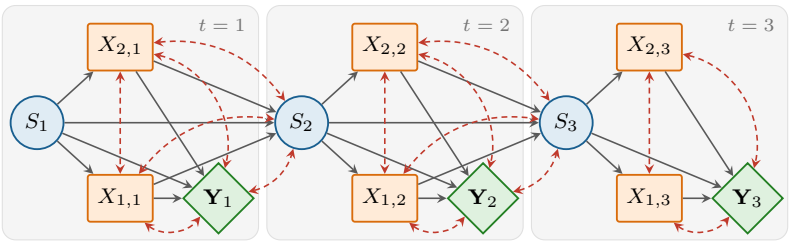


Figure 2: Causal diagram of a Causal Markov Game with $n = 2$ agents over $H = 3$ stages

A joint counterfactual policy is decentralized by construction: agent i's decision rule $\pi_{i,t}$ reads only its own intuition $X'_{i,t}$, never $X'_{j,t}$ for $j \neq i$: a restriction made natural by (3). The L_1 and L_2 policy spaces from Sec. 1.1 are recovered as boundary cases: setting $\pi_{i,t}(X_{i,t} \mid s, x') = \delta_{X_{i,t}=x'}$ recovers an L_1 policy, while letting $\pi_{i,t}(\cdot \mid s, \cdot)$ be constant in its second argument recovers an L_2 policy. Hence $\Pi_i^{(1)} \cup \Pi_i^{(2)} \subseteq \Pi_i^{\mathrm{CTF}}$, with the inclusion strict in general.

Example 2.3 (ICPD as a Causal Markov Game)**.** ICPD is the CMG with $n = 2$, $\mathcal{A}_i = \{0, 1\}$, exogenous variables $U_{i,t} \in \{0, 1\}$ with $P(U_{i,t} = 0) = 0.6$ independently across i and t, degenerate reward noise, a trivial state, $f_{X_i}(S_t, U_{i,t}) = R_{i,t} \oplus U_{i,t}$, and f_{Y_i} given by the payoff structure in Fig. 1a. Environments M_1 and M_2 correspond to $R_{i,t} = 1$ and $R_{i,t} = 0$, respectively, fixing different f_{X_i} in the same family. U_t^C is trivial and the $U_{i,t}$ are independent, so (3) holds given S_t alone.

L_1 payoffs in M_2. In M_2, $X_{i,t} = U_{i,t}$, so each agent defects with probability 0.4 and cooperates with probability 0.6 under its natural mechanism. The expected per-round payoff is

$$\mathbb{E}[Y_{i,t}] \;=\; \sum_{u_1,u_2} P(u_1)P(u_2)\, f_{Y_i}(u_1, u_2,\, u_1, u_2) \;=\; -2.424, \tag{5}$$

where the sum aggregates the four joint (u_1, u_2) outcomes weighted by the payoff matrix in Fig. 1a. Both agents receive -2.424, giving $\mu_1 = (-2.424, -2.424)$.

L_2 Nash payoff in M_1 and M_2. For any deterministic joint L_2 policy $\boldsymbol{\sigma}$, the intervention $\mathrm{do}(\mathbf{X}_t = \boldsymbol{\sigma}(S_t))$ severs f_{X_i} in both M_1 and M_2. Since f_{Y_i} and $P(\mathbf{U})$ are identical across the two environments, the interventional reward distribution coincides in M_1 and M_2

$$P_{\boldsymbol{\sigma}}(Y_{i,t} \mid s_t) \;=\; \sum_{u_1,u_2} P(u_1)P(u_2)\, \mathbb{1}\{Y_{i,t} = f_{Y_i}(\sigma_1(s_t), \sigma_2(s_t),\, u_1, u_2)\}. \tag{6}$$

The dominant strategy in either environment is always-defect, with Nash payoff $\mu_{\mathrm{NE}} = (-1.92, -1.92)$.

Counterfactual Payoff in M_2. Consider the counterfactual policy $\pi_{i,t}^{\mathrm{opp}}(X_{i,t} = 1 - x' \mid S_t, x') = 1$, which acts opposite to its intuition. In M_2, the intuition $X'_{i,t} = U_{i,t}$ is systematically inverted relative to circumstances; π^{opp} undoes this inversion, producing $X_{i,t} = 1 - U_{i,t}$ — i.e., the same realized action as the natural mechanism in M_1. The expected per-round payoff becomes

$$\mathbb{E}_{\pi^{\mathrm{opp}}}[Y_{i,t}] \;=\; \sum_{u_1,u_2} P(u_1)P(u_2)\, f_{Y_i}(1 - u_1, 1 - u_2,\, u_1, u_2) \;=\; 0, \tag{7}$$

giving the joint payoff $\mu_2 = (0, 0)$. The same policy applied in M_1 would invert the agents' favorable intuitions and yield the dominated $\mu_1 = (-2.424, -2.424)$. This illustrates that the optimal counterfactual policy depends on causal structure that the interventional projection cannot detect. □

The examples above are not specific to the Iterated Prisoner's Dilemma. The next theorem shows that for *every* Markov game one can construct a pair of CMGs with the same interventional projection but with L_1 payoffs lying on opposite sides of the Markov-game Nash payoff in the Pareto order.

Theorem 2.4 (CMG strictly generalizes MG)**.** *Let G be any finite-horizon Markov game with horizon H and finite state and action spaces, and $|\mathcal{A}_i| \geq 2$ for some agent i. Fix a Markov Nash Equilibrium of G, with expected total payoff vector μ_{NE} from the initial state. There exist Causal Markov Games*

$\mathcal{M}_1$ and $\mathcal{M}_2$, with the same interventional transition kernels and expected reward functions as G, whose L_1 policy payoff vectors μ_1, μ_2 satisfy

$$\mu_1 \prec \mu_{\mathrm{NE}} \prec \mu_2. \tag{8}$$

Here $\prec$ denotes the strict Pareto order.

The proof is provided in Appendix C.1, where we provide a construction with strict separation and arbitrarily large gaps when reward magnitudes are unrestricted. With rewards bounded in $[0, 1]$, the gap can grow linearly in H (Appendix C.2 gives a two-player instance). This result makes precise the sense in which a Markov game is causally underdetermined: structurally distinct CMGs share an interventional projection while their behavioral and counterfactual payoffs lie on either side of μ_{NE}. When intuitions encode useful structure, counterfactual policies can reach payoffs unattainable through interventional reasoning alone, motivating the need for learning algorithms studied in Sec. 3.

3 Learning in Causal Markov Games

We now develop algorithms for learning from samples under progressively weaker informational assumptions: in Sec. 3.1, we present an algorithm for the setting where other agents' natural intuitions are observable after each step; in Sec. 3.2, we relax this assumption and propose an algorithm for two-player games; in Sec. 3.3, we extend deep MARL architectures to counterfactual action spaces.

3.1 Learning with Observable Natural Actions

We work in the tabular episodic regime with finite state space $\mathcal{S}$ ($|\mathcal{S}| = S$), action domains $\mathcal{A}_i$ of size A_i, horizon H, and n agents. Throughout this section, rewards satisfy $0 \le Y_{i,t} \le 1$ almost surely. Rewards and transitions both depend on U_t, and are hence confounded. However, given $(S_t, \mathbf{X}'_t)$, the executed actions carry no further information about U_t. $(S_t, \mathbf{X}'_t)$ is a Markov state (Lem. B.3), with intuitions at state s following a joint distribution $\rho_t(\mathbf{x}' \mid s) := P(\mathbf{X}'_t = \mathbf{x}' \mid S_t = s)$. The coordinates of $\mathbf{X}'_t$ may be correlated through U_t^C. Note that each agent only sees its own intuition before acting and nothing of the others'. In this section, we assume *post-hoc observability*, that is, at the end of each step the learner records the full tuple $(S_t, \mathbf{X}'_t, \mathbf{X}_t, \mathbf{Y}_t, S_{t+1})$, including every agent's intuition. Opponents' intuitions become available to the learner only after the agents have acted, and the policies depend on the state and the agent's own intuition. This observation assumption is compatible with centralized training and decentralized execution (CTDE) when natural actions can be logged. The goal of learning is to find an approximate *Causal Nash Equilibrium* (CNE), a product counterfactual policy profile from which no agent can improve by deviating unilaterally.

For a joint counterfactual policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ and initial state s_1, let $W^{\boldsymbol{\pi}}_{1,i}(s_1)$ denote the expected total reward of agent i over the episode: the expectation is over the natural actions $\mathbf{X}'_t$ drawn by the environment from $\rho_t(\cdot \mid S_t)$ at each step, the randomness in $\boldsymbol{\pi}$, and the state transitions. We write $\boldsymbol{\pi}_{-i} = (\pi_j)_{j \ne i}$ for the joint policy of all agents except i.

Definition 3.1 (ϵ-Approximate CNE)**.** A joint counterfactual policy $\boldsymbol{\pi} \in \prod_{i=1}^n \Pi_i^{\mathrm{CTF}}$ is an *ϵ-approximate Causal Nash Equilibrium* (CNE) of the CMG $\mathcal{M}$ if no agent can improve its expected total reward by more than ϵ through a unilateral deviation to any counterfactual policy:

$$\max_{i \in [n]} \sup_{\pi'_i \in \Pi_i^{\mathrm{CTF}}} \left(W_{1,i}^{(\pi'_i, \boldsymbol{\pi}_{-i})}(s_1) \; - \; W^{\boldsymbol{\pi}}_{1,i}(s_1) \right) \; \le \; \epsilon. \tag{9}$$

The key distinction from a standard Nash equilibrium is that deviations are restricted to counterfactual policies, and agent i conditions on $(S_t, X'_{i,t})$, not on other agents' intuitions. Throughout the paper we use "L_3 policy" and "counterfactual policy" interchangeably; both refer to elements of Π_i^{CTF}.

Algorithm: CNash-VI-FO. We propose *Causal Nash Value Iteration with post hoc observability of natural intuitions* (Alg. 2), a model-based optimistic algorithm that learns a CNE from episodic interactions. It operates on the *augmented state* $\tilde{s}_t = (s_t, x'_{1,t}, \ldots, x'_{n,t})$, which appends each agent's natural action to the environment state, and runs optimistic backward induction on the empirical augmented model.

The critical step is the BAYESNASH subroutine: at each state s, the agent can only act on its own intuition and the shared state. So, it reduces to a Bayesian Game, and this subroutine computes a

Bayesian Nash equilibrium (BNE) of the game where each agent's type is its natural action x'_i, drawn from the empirical joint distribution $\hat{\rho}_t(\cdot \mid s) \approx \rho_t(\cdot \mid s)$. An agent of type x'_i evaluates its actions under the interim belief $\hat{\rho}_t(x'_{-i} \mid s, x'_i)$. When the intuitions share a confounder, an agent's own intuition tells it something about the others' intuitions, and the equilibrium uses that information through this belief. Since each agent in this Bayesian Game conditions only on its own type (s, x'_i), the resulting equilibrium is a product counterfactual policy by construction. The algorithm returns the policy minimizing the optimism gap $\Delta_k = \max_i\big(\overline{W}^k_{1,i}(s_1) - \underline{W}^k_{1,i}(s_1)\big)$ over episodes k; we denote its gap by $\Delta_{\text{out}} = \min_{1\le k\le K} \Delta_k$, where $\overline{W}_{1,i}$ and $\underline{W}_{1,i}$ are optimistic and pessimistic estimates of the expected total reward $W_{1,i}$. CNash-VI-FO estimates the state transition probabilities and the joint intuition distribution separately. The algorithm and its parameter settings are in Appendix D.4.

Theorem 3.2 (Sample Complexity of CNash-VI-FO)**.** *There exists an absolute constant $C > 0$ such that, for any $p \in (0,1)$, with probability at least $1-p$, CNash-VI-FO (Alg. 2) outputs an ϵ-approximate Causal Nash Equilibrium provided the number of episodes satisfies*

$$K \;\ge\; C\,\frac{H^4\,S^2\,\big(\prod_i A_i\big)^2\,\iota}{\epsilon^2}, \tag{10}$$

where $\iota = \log\big(S(\prod_i A_i)^2 KH/p\big)$.

This upper bound has an additional factor of $\prod_i A_i$ compared with the $\widetilde{O}(H^4S^2\prod_i A_i/\epsilon^2)$ bound for standard n-player Markov games [Liu et al., 2021]. The proof is in Appendix D.6.

3.2 Learning without Observable Natural Actions

We now consider a more realistic setting where the opponent's natural action is *never* observed: an agent sees its own intuition $X'_{1,t}$ and the opponent's executed action $X_{2,t}$, but not $X'_{2,t}$. Without post-hoc observability the augmented state cannot be formed, preventing the direct conditional estimation used in Section 3.1. For two-player general-sum CMGs, we exploit a key observation: under a uniform exploration policy, observed rewards and transitions are finite mixtures over the opponent's natural intuitions and the conditional model can be recovered under certain identification assumptions (Appendix E.1). Here, we restrict our attention to two players ($n = 2$). Let x'_i denote the natural intuition and x_i the executed action, and write $a_i := (x'_i, x_i) \in \mathcal{A}_i \times \mathcal{A}_i$ for the pair. The conditional reward and transition distributions are indexed by (s, a_1, a_2), where $a_i = (x'_i, x_i)$; from agent 1's perspective, x'_2 is unobserved.

Algorithm: Causal Nash VI without Observability of Natural Intuitions (CNash-VI-NO). The algorithm proceeds in three phases: exploration, identification, and exploitation.

Phase 1: Exploration. For K_{exp} episodes, both agents play uniformly at random over their executed actions, using independent randomization and ignoring their own natural intuitions. Agent 1 records $(s_t, a_{1,t}, x_{2,t}, r_t, s_{t+1})$ at each step, where $a_{1,t} = (x'_{1,t}, x_{1,t})$ and $r_t = (Y_{1,t}, Y_{2,t})$ is the observed reward pair. Let R_t denote the conditional distribution of this reward pair. For each (s, x'_1) of positive probability, define the unknown mixing weights $\rho_{2,t}(x'_2 \mid s, x'_1) := P(X'_{2,t} = x'_2 \mid S_t = s, X'_{1,t} = x'_1)$. Under uniform exploration, the executed actions are independent of the joint intuition conditional on the state. Thus, marginalizing over the opponent's unobserved natural intuition gives

$$R_t(\cdot \mid s, a_1, x_2) = \sum_{x'_2 \in \mathcal{A}_2} \rho_{2,t}(x'_2 \mid s, x'_1)\, R_t(\cdot \mid s, a_1, x_2, x'_2), \tag{11}$$

where $R_t(\cdot \mid s, a_1, x_2, x'_2)$ are the unknown component reward distributions. The mixing weights are invariant in the executed actions (x_1, x_2) at fixed (s, t, x'_1). An analogous mixture holds for transitions, for $t < H$:

$$P_t(s' \mid s, a_1, x_2) = \sum_{x'_2 \in \mathcal{A}_2} \rho_{2,t}(x'_2 \mid s, x'_1) P_t(s' \mid s, a_1, x_2, x'_2). \tag{12}$$

Phase 2: Identification. We impose Assumptions E.1–E.3 in Appendix E.1. The reward-family and distinctness conditions identify the reward components and their weights at informative cells up to permutation. Assumption E.3 additionally requires identification procedures IDENT_r and IDENT_P

Algorithm 1 CNash-VI-NO

Require: CMG $\mathcal{M}$ (two-player), horizon H, exploration episodes K_{exp}, identification oracles IDENT_r, IDENT_P.

Phase 1: Exploration
for episode $k = 1, \ldots, K_{\text{exp}}$ **do**
  Start an independent episode at s_1.
  for step $t = 1, \ldots, H$ **do**
    Observe state s_t and own natural intuition $x'_{1,t}$.
    Both agents independently play $x_{i,t} \sim \text{Unif}(\mathcal{A}_i)$, $i = 1, 2$; agent 1 observes $x_{2,t}$, rewards $(r_{1,t}, r_{2,t})$, and next state s_{t+1}.
    Set $a_{1,t} := (x'_{1,t}, x_{1,t})$ and store $(s_t, a_{1,t}, x_{2,t}, r_{1,t}, r_{2,t}, s_{t+1})$.
  end for
end for
Phase 2: Identification
The identification calls below use the full exploration data and the identification information required by Assumption E.3.
for each observed $(s, t) \in \mathcal{S} \times [H]$ **do**
  Estimate $\hat{\rho}_{1,t}(\cdot \mid s)$ from observed own natural intuitions at state s, step t.
  Recover $\hat{\rho}_{2,t}(\cdot \mid s, x'_1)$ for each observed x'_1 using the identification procedures of Assumption E.3.
  for each observed (a_1, x_2) at (s, t) **do**
    $\hat{r}_t(s, a_1, x_2, \cdot) \leftarrow \text{IDENT}_r\big(\text{exploration data};\ s, a_1, x_2, t\big)$.
    if $t < H$ **then**
      $\hat{P}_t(\cdot \mid s, a_1, x_2, \cdot) \leftarrow \text{IDENT}_P\big(\text{joint reward–next-state data};\ s, a_1, x_2, t\big)$.
    end if
  end for
  Align the opponent-type labels across action cells and own-intuition values, and convert them to actual intuition labels as in Assumption E.3.
end for
Use reward estimates in $[0, 1]$ and valid probability distributions, with arbitrary valid defaults at unobserved rows.
Set $\hat{\rho}_t(x'_1, x'_2 \mid s) = \hat{\rho}_{1,t}(x'_1 \mid s)\hat{\rho}_{2,t}(x'_2 \mid s, x'_1)$ for every (s, t).
Phase 3: Exploitation
Construct augmented model $\hat{\tilde{\mathcal{M}}}$ with $\tilde{\mathcal{S}} = \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$, using $\hat{r}$, $\hat{P}$, and $\hat{\rho}$.
$\hat{\boldsymbol{\pi}} \leftarrow$ Solve CNE of $\hat{\tilde{\mathcal{M}}}$ by exact backward induction with BAYESNASH and joint prior $\hat{\rho}_t(\cdot \mid s)$, as in Definition D.2.
return $\hat{\boldsymbol{\pi}}$.

that consistently recover both players' conditional reward means, the transition probabilities, and the conditional weights $\rho_{2,t}(\cdot \mid s, x'_1)$. The recovered type labels must agree across executed-action cells and own-intuition values, and their correspondence to the agents' observed intuitions must be available when executing the policies. Reward and next state may remain dependent conditional on the state, intuitions, and executed actions.

Phase 3: Exploitation. Let $\hat{\rho}_{1,t}(\cdot \mid s)$ be the empirical distribution of agent 1's observed intuitions and let $\hat{\rho}_{2,t}(\cdot \mid s, x'_1)$ be the identified conditional weights. Form the estimated joint intuition distribution

$$\hat{\rho}_t(x'_1, x'_2 \mid s) = \hat{\rho}_{1,t}(x'_1 \mid s)\hat{\rho}_{2,t}(x'_2 \mid s, x'_1).$$

With the estimated conditional model $\hat{M} = (\hat{r}, \hat{P}, \hat{\rho})$, compute an exact equilibrium by backward induction using the Bayesian stage games of Definition D.2 with joint prior $\hat{\rho}_t(\cdot \mid s)$. The output is a product counterfactual policy $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2)$, where each $\hat{\pi}_{i,t}(\cdot \mid s, x'_i)$ depends only on the state and that player's own intuition.

This extends the learning algorithm of Maiti et al. [2025] from single-stage causal games to the episodic Markov setting.

Theorem 3.3 (Asymptotic Convergence of CNash-VI-NO)**.** *Consider a two-player general-sum CMG with finite state and action spaces, finite horizon H, and rewards in $[0, 1]$ satisfying Assumptions E.1–*

E.3, with exact Bayesian equilibrium planning. Let $\hat{\boldsymbol{\pi}}_K$ be the output of Algorithm 1 after $K_{\text{exp}} = K$ exploration episodes. Then

$$\max_{i\in\{1,2\}} \sup_{\pi'_i\in\Pi_i^{\text{CTF}}} \left(W_{1,i}^{(\pi'_i,\hat{\boldsymbol{\pi}}_{K,-i})}(s_1) - W_{1,i}^{\hat{\boldsymbol{\pi}}_K}(s_1)\right) \xrightarrow{\text{a.s.}} 0 \quad \textit{as } K\to\infty, \tag{13}$$

i.e., the exploitability of $\hat{\boldsymbol{\pi}}_K$ vanishes almost surely and any limit point is a Causal Nash Equilibrium.

The proof is provided in Appendix E. Unlike the simultaneous explore-exploit approach of Sec. 3.1, this algorithm separates exploration from exploitation, using additional identification assumptions to operate without observing the opponent's natural actions.

3.3 Deep MARL with Counterfactual Actions

The tabular algorithms of Secs. 3.1 and 3.2 do not scale to high-dimensional observations or large action spaces. In this section, we extend deep MARL to the counterfactual action space: an L_3 agent receives its natural action $x'_{i,t}$ as an additional input, treating it as a privately observed Bayesian type. The opponents' interface is unchanged from each opponent's perspective; the agent remains part of the environment.

Causal MADDPG: We instantiate this extension on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [Lowe et al., 2017]. Standard MADDPG trains, for each agent i, a local actor $\pi_{\theta_i}(o_i)$ against a centralized critic $Q_{\phi_i}(s,\mathbf{x})$. Causal MADDPG conditions both on the natural-action context: $\pi_{\theta_i}(o_i,x'_i)$ and $Q_{\phi_i}(s,\mathbf{x}',\mathbf{x})$. Each actor learns a type-conditioned response, the deep analogue of the per-type BNE strategy $\pi_{i,t}(\cdot\mid s,x'_i)$ from Sec. 3.1. The critic conditions on the opponents' logged intuitions $\mathbf{x}'_{-i}$ rather than marginalizing over them, lowering the variance of the deterministic policy gradient, at the cost of a richer training-time input. Conditioning the critic on $\mathbf{x}'$ aligns the training signal with the CMG, and restricting the actor to (o_i,x'_i) rather than the full $(o_i,\mathbf{x}')$ enforces the L_3 restriction by construction. The same idea can be applied to any CTDE method with local actors and a centralized critic, such as MAPPO [Yu et al., 2022] and HAPPO [Kuba et al., 2022]. We show our evaluation of Causal MADDPG against the standard version in Sec. 4, with more details in Appendix F.3.

4 Experiments

In this section, we validate (i) CNash-VI-FO (Sec. 3.1) on three tabular CMGs, where we examine the qualitative convergence behavior suggested by Thm. 3.2; (ii) CNash-VI-NO (Sec. 3.2) on the Iterated Causal Prisoner's Dilemma; and (iii) the deep MARL augmentation (Sec. 3.3) on a confounded Multi-Particle Environment (MPE).

4.1 CNash-VI-FO on Tabular CMGs

Thm. 3.2 bounds the gap Δ_{out} by $\widetilde{O}(K^{-1/2})$ and uses it as an upper bound on exploitability (App. D.6). We evaluate an empirical variant of Alg. 2 on three instances: (1) the ICPD of Ex. 1.1 in M_1 (H=5); (2) a random CMG with state transitions and independent intuitions (S=3, A_1=A_2=2, H=4); and (3) the same CMG as in (2) with correlated intuitions. Over 5 seeds and $K \le 2\cdot 10^5$ episodes, the log-log slope of Δ_{out} against K is -0.52, -0.58 and -0.57 on the three instances ($R^2 \ge 0.97$), against the reference slope -0.50 (Fig. 3). Exploitability stayed below Δ_{out} at all 180 tested seed-checkpoint pairs and is below 0.004 for the two random CMGs; on the ICPD it is exactly 0 from $K \approx 6\cdot 10^4$ on. The theorem gives the same upper bound with correlated intuitions; the observed decay is similar across the two random instances. More details are in App. F.1.

4.2 Windy Multi-Particle Environment

We augment the classic MPE with the wind described below; the natural-action input is computed by the simulator from the true state (App. F.2). These partially observed experiments evaluate the counterfactual-input architecture; the tabular guarantees do not directly apply to this setting.

Environment Setup. The PettingZoo MPE task: `simple_spread` (cooperative) [Lowe et al., 2017, Terry et al., 2021] is extended with a wind state machine that pushes every target and blacks out each

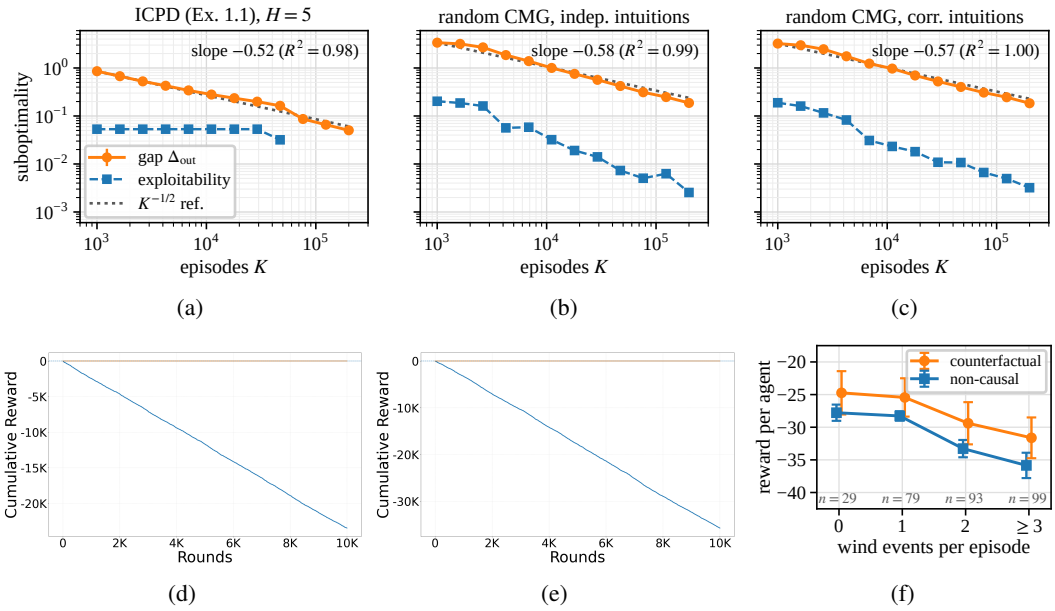


Figure 3: (a–c) Empirical CNash-VI-FO on the ICPD, a random CMG, and the same CMG with correlated intuitions: gap Δ_{out} (mean $\pm$ std over 5 seeds), exact exploitability, and a $K^{-1/2}$ reference, against episodes K. (d, e) Cumulative reward over 10^4 ICPD rounds against always-defect and tit-for-tat. (f) Per-agent reward against wind events on `simple_spread` (mean $\pm$ std over 5 seeds; n = evaluation episodes per bin). Orange: counterfactual; blue: non-causal.

agent's visible observation slots (15–24% of steps). Additionally, the wind also pushes the agent in its direction, and *counterfactual* agents can act based on the natural action $X'_{i,t}$; *non-causal* agents cannot. The agents are trained with standard and Causal MADDPG over 5 seeds and evaluated on 300 matched-seed episodes; environment setup, algorithm, and reproducibility details are provided in Appendix F. In Fig. 3f, the aggregate per-agent gain with counterfactual actions is $+3.6 \pm 3.0$ (mean $\pm$ std), and the gap is largest in the two windiest cases ($+3.9$ and $+4.2$).

4.3 Iterated Causal Prisoner's Dilemma

We run CNash-VI-NO on the ICPD of Ex. 1.1 with 10^5 rounds of uniform-random exploration, identification, and then backward-induction exploitation. Deployed for 10^4 rounds against always-defect (D) and tit-for-tat (TfT), the learned policy stays near zero while its opponent accumulates about -24K and -35K respectively (Fig. 3d-e). In the stage game, the optimal counterfactual policy earns strictly greater expected reward than every interventional strategy of agent 1 against every interventional strategy of agent 2. This advantage also holds in expected cumulative reward against history-dependent interventional opponents such as win-stay-lose-shift, generous TfT, and no-regret learners (Appendix F.5).

5 Conclusions

Causal Markov Games lift multi-agent sequential decision-making to the counterfactual layer of the PCH and strictly generalize Markov Games in the presence of unobserved confounders. We develop a sample-efficient learner under post-hoc observability of natural actions, an explore-then-exploit algorithm with asymptotic guarantees when opponents' natural actions are never observed, and counterfactual augmentation for deep MARL. Empirically, counterfactual agents have an advantage in the ICPD and higher mean reward on windy `simple_spread`. We hope CMGs provide a foundation for future work on causal multi-agent learning.

Acknowledgments and Disclosure of Funding

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

E. Bareinboim, J. Zhang, and S. Lee. An introduction to causal reinforcement learning. Technical Report R-65, Causal Artificial Intelligence Lab, Columbia University, Dec 2024. https://causalai.net/r65.pdf.

Elias Bareinboim. Causal artificial intelligence: A roadmap for building causally intelligent systems. 2025. URL https://causalai-book.com/. Book draft, October 11, 2025.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33: 5527–5540, 2020.

Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th National Conference on Artificial Intelligence*, AAAI'04, page 709–715. AAAI Press, 2004. ISBN 0262511835.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EcGGFkNTxdJ.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Michael L. Littman. Friend-or-foe q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 322–328, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

A. Maiti, P. Jain, and E. Bareinboim. Counterfactual rationality: A causal approach to game theory. Technical Report R-125, Causal Artificial Intelligence Lab, Columbia University, USA, January 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

Jordan Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Junzhe Zhang and Elias Bareinboim. Can humans be out of the loop? In *Conference on Causal Learning and Reasoning*, pages 1010–1025. PMLR, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: References to the contribution points have been added to abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Assumptions and limitations have been discussed wherever applicable.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The proofs have been provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All algorithmic and experimental details are discussed in the Appendix as well as the main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Link to the code repository is provided in the Appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details as well as link to the code are provided in the main paper and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: They are discussed in the experiment section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources for this work is trivial and is run on a Macbook Pro.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work studies fundamental models and algorithms and does not evaluate a specific societal deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose such a risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets, except standard programming languages and libraries.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not involve human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not an important, original, or non-standard component of the core methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Supplementary Material

A Notations **20**

B The Counterfactual Submodel **20**

B.1 Definition 21

B.2 Markov Property of the Counterfactual Submodel 22

C Proofs for Section 2 **22**

C.1 Theorem 2.4: CMG strictly generalizes MG 22

C.2 Proposition C.3: Separation under Bounded Rewards 23

D Proofs for Section 3.1 **24**

D.1 Conditional Model 25

D.2 Value Functions 25

D.3 The Bayesian Equilibrium Oracle 25

D.4 CNash-VI-FO with Factored Estimation 26

D.5 The Confidence Gap Bounds Exploitability 27

D.6 Proof of Theorem 3.2 28

D.7 Computational Cost of the Bayesian Equilibrium Oracle 29

D.8 Role of Post-hoc Observability 30

E Proofs for Section 3.2 **31**

E.1 Identifiability Conditions 31

E.2 Proof of Theorem 3.3 32

E.3 Discussion of Identifiability 33

E.4 A Finite-Sample Rate in the Atomic Case 33

F Experiments **34**

F.1 CNash-VI-FO on Tabular CMGs 34

F.2 Environment: Windy Multi-Particle Environment 35

F.3 Algorithm: MADDPG with Counterfactual Inputs 36

F.4 Full Results 36

F.5 Iterated Causal Prisoner's Dilemma 37

A Notations

Table S1 presents a glossary of the notations used in the paper.

B The Counterfactual Submodel

This appendix introduces the counterfactual submodel induced by a joint counterfactual policy and establishes its Markov property.

Table S1: Notations.

Symbol	Meaning
n, H, K	number of agents, horizon, number of episodes
$S_t \in \mathcal{S}$	state at stage t; $\lvert\mathcal{S}\rvert = S$
$X_{i,t} \in \mathcal{A}_i$	executed action of agent i; $\lvert\mathcal{A}_i\rvert = A_i$
$X'_{i,t} = f_{X_i}(S_t, U_{i,t}, U_t^C)$	natural action (intuition) of agent i; observed by agent i before acting; logged for all players post hoc in FO
$Y_{i,t}$	reward of agent i
$U_{i,t}, U_{y_i,t}, U_{s,t}, U_t^C$	agent, reward, state and shared latents at stage t; observed by no one
L_1, L_2, L_3	observational (natural), interventional and counterfactual layers of the PCH
$\Pi_i^{(1)}$, $\Pi_i^{(2)}$, Π_i^{CTF}	natural, interventional (Sec. 1.1) and counterfactual (Def. 2.2) policy classes
$\rho_{i,t}(x' \mid s)$	$P(X'_{i,t} = x' \mid S_t = s)$, marginal distribution of agent i's natural action (Sec. 3.1)
$\rho_t(\mathbf{x}' \mid s)$	$P(\mathbf{X}'_t = \mathbf{x}' \mid S_t = s)$, joint distribution of the natural actions (Sec. 3.1)
$\rho_{2,t}(x'_2 \mid s, x'_1)$	conditional distribution of the opponent's natural action (Sec. 3.2)
$\tilde{s}_t = (s_t, \mathbf{x}'_t)$	augmented state: environment state plus all natural actions (Sec. 3.1); $\tilde{S} = \lvert\tilde{\mathcal{S}}\rvert = S\prod_i A_i$
$W_{1,i}^{\boldsymbol{\pi}}(s_1)$	expected total reward of agent i under $\boldsymbol{\pi}$ from s_1 (Sec. 3.1)
FO / NO	natural actions of all agents are logged post hoc (Sec. 3.1) / the opponent's never is (Sec. 3.2)
CNE	Causal Nash Equilibrium (Def. 3.1)

We use finite action domains, standard Borel variable spaces, and measurable structural functions. Fix $S_1 = s_1$; a random initial state may instead be generated from the independent state noise $U_{s,1}$. Stage indices on structural functions are displayed explicitly.

B.1 Definition

Following a counterfactual policy in the original CMG produces a new SCM in which the agent's intuition becomes an explicit endogenous variable.

Definition B.1 (Counterfactual Submodel)**.** Let $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U})\rangle$ be a CMG (Def. 2.1) and $\boldsymbol{\pi}$ a joint counterfactual policy (Def. 2.2). Introduce $Z_{i,t} \sim \mathrm{Unif}[0,1]$, mutually independent and independent of $\mathbf{U}$. Choose sampling maps $g_{i,t}^{\boldsymbol{\pi}}$ satisfying

$$P_Z\big(g_{i,t}^{\boldsymbol{\pi}}(s, x'_i, Z) = x_i\big) = \pi_{i,t}(x_i \mid s, x'_i).$$

These maps exist by inverse-transform sampling on the finite action domains.

For terminal-state notation, include S_{H+1} and a fresh independent state noise $U_{s,H+1}$ with the prescribed state-noise distribution, if they are not already present. The *counterfactual submodel* is

$$\mathcal{M}_{\boldsymbol{\pi}} = \big\langle \mathbf{U}_{\boldsymbol{\pi}}, \mathbf{V}_{\boldsymbol{\pi}}, \mathcal{F}_{\boldsymbol{\pi}}, P_{\boldsymbol{\pi}} \big\rangle, \tag{14}$$

where $\mathbf{U}_{\boldsymbol{\pi}}$ augments $\mathbf{U}$ by the new noises, $\mathbf{V}_{\boldsymbol{\pi}} = \mathbf{V} \cup \mathbf{X}' \cup \{S_{H+1}\}$, and $P_{\boldsymbol{\pi}}$ extends $P(\mathbf{U})$ by the corresponding independent product factors. Here $\mathbf{X}' = \{X'_{i,t} : i \in [n],\ t \in [H]\}$ are the agents' intuitions. Starting from $S_1 = s_1$, the structural assignments are

$$\begin{aligned} X'_{i,t} &\leftarrow f_{X_i,t}(S_t, U_{i,t}, U_t^C), \\ X_{i,t} &\leftarrow g_{i,t}^{\boldsymbol{\pi}}(S_t, X'_{i,t}, Z_{i,t}), \\ Y_{i,t} &\leftarrow f_{Y_i,t}(S_t, \mathbf{X}_t, \mathbf{U}_{a,t}, U_t^C, U_{y_i,t}), \\ S_{t+1} &\leftarrow f_{S,t}(S_t, \mathbf{X}_t, \mathbf{U}_{a,t}, U_t^C, U_{s,t+1}), \end{aligned} \tag{15}$$

for $t \in [H]$ and $i \in [n]$. There are no actions or rewards after stage H.

The submodel makes $\mathbf{X}'$ an explicit mediator between the unobserved variables and the realized actions. Each agent's policy still uses only its own intuition. The following Markov property is proved in App. B.2.

Remark B.2 (Markov property of the submodel). Let $\mathcal{H}_t = \sigma(S_{1:t}, \mathbf{X}'_{1:t}, \mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})$. In $\mathcal{M}_{\boldsymbol{\pi}}$, the augmented state $\widetilde{S}_t = (S_t, \mathbf{X}'_t)$ satisfies, for $t < H$,

$$\begin{aligned} &P(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \mid \mathcal{H}_t) \\ &\quad = P(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \mid S_t, \mathbf{X}'_t, \mathbf{X}_t) \end{aligned} \tag{16}$$

almost surely, where P denotes the corresponding conditional probability distribution under $\mathcal{M}_{\boldsymbol{\pi}}$. Current reward and next state may remain dependent given $(S_t, \mathbf{X}'_t, \mathbf{X}_t)$, so the proof keeps their joint conditional distribution.

B.2 Markov Property of the Counterfactual Submodel

Lemma B.3 (Markov property of $\mathcal{M}_{\boldsymbol{\pi}}$)**.** *Let $\mathcal{M}$ be a Causal Markov Game (Def. 2.1) and $\boldsymbol{\pi}$ a joint counterfactual policy (Def. 2.2). Equation* (16) *holds almost surely for every $t < H$. At $t = H$, the same assertion holds for $(S_{H+1}, \mathbf{Y}_H)$. The conditional distributions can be chosen independently of $\boldsymbol{\pi}$.*

Proof. By independence of the exogenous variables, $(\mathbf{U}_{a,t}, U_t^C)$ is independent of the past and S_t. The current intuitions depend only on S_t and these latents, while the executed actions use independent policy randomization. Hence

$$P(\mathbf{U}_{a,t}, U_t^C \mid \mathcal{H}_t) = P(\mathbf{U}_{a,t}, U_t^C \mid S_t, \mathbf{X}'_t).$$

For $t < H$, the remaining noises $(\mathbf{U}_{y,t}, U_{s,t+1}, \mathbf{U}_{a,t+1}, U_{t+1}^C)$ are jointly independent of $(\mathcal{H}_t, \mathbf{U}_{a,t}, U_t^C)$. The structural equations therefore imply that the *joint* conditional distribution of $(S_{t+1}, \mathbf{Y}_t, \mathbf{X}'_{t+1})$ depends on the history only through $(S_t, \mathbf{X}'_t, \mathbf{X}_t)$. This proves (16).

In particular, its marginals and the independence of the next-stage intuition latents give

$$P(\mathbf{Y}_t \mid \mathcal{H}_t) = P(\mathbf{Y}_t \mid S_t, \mathbf{X}'_t, \mathbf{X}_t), \tag{17}$$

$$P(S_{t+1} \mid \mathcal{H}_t) = P(S_{t+1} \mid S_t, \mathbf{X}'_t, \mathbf{X}_t), \tag{18}$$

$$P(\mathbf{X}'_{t+1} \mid S_{t+1}, \mathbf{Y}_t, \mathcal{H}_t) = P(\mathbf{X}'_{t+1} \mid S_{t+1}). \tag{19}$$

These distributions are determined by the structural equations and exogenous distributions, with a fixed choice on zero-probability conditioning events, so they can be chosen independently of $\boldsymbol{\pi}$. At $t = H$, the same argument applies after omitting $\mathbf{X}'_{H+1}$ and its latents. □

C Proofs for Section 2

Let G be the finite Markov game in Theorem 2.4, with n players, horizon $H \geq 1$, initial state s_1, state space $\mathcal{S}$, and action sets $\mathcal{A}_1, \ldots, \mathcal{A}_n$. For a joint action $\mathbf{x} = (x_1, \ldots, x_n)$, write $P_t(s' \mid s, \mathbf{x})$ for the probability of moving from s to s' at stage t, and $r_{j,t}(s, \mathbf{x}) \in \mathbb{R}$ for player j's expected reward at that step. All payoffs below are expected total rewards over the H stages, starting from s_1.

Let $\mathrm{do}(\mathbf{x})$ denote the joint intervention $\mathrm{do}(\mathbf{X}_t = \mathbf{x})$ at stage t. A CMG $\mathcal{M}$ with the same horizon, initial state, and state/action spaces has interventional projection G if, for every $t, s, s', \mathbf{x}, j$,

$$\begin{aligned} P(S_{t+1} = s' \mid S_t = s, \mathrm{do}(\mathbf{x}); \mathcal{M}) &= P_t(s' \mid s, \mathbf{x}), \\ \mathbb{E}[Y_{j,t} \mid S_t = s, \mathrm{do}(\mathbf{x}); \mathcal{M}] &= r_{j,t}(s, \mathbf{x}). \end{aligned} \tag{20}$$

Here the probabilities and expectations are evaluated in $\mathcal{M}$ using its structural equations at the displayed state and action. This also defines them at states not reached during play. For vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, write $\mathbf{v} \prec \mathbf{w}$ when $v_j \leq w_j$ for every j, with at least one strict inequality.

C.1 Theorem 2.4: CMG strictly generalizes MG

Proof of Theorem 2.4. Fix a Markov NE of G, and let $\boldsymbol{\mu}_{\mathrm{NE}} = (\mu_{\mathrm{NE},j})_{j=1}^n$ be its payoff vector.

Construction. Choose an agent i with $|\mathcal{A}_i| \geq 2$, distinct $a, b \in \mathcal{A}_i$, and one action $c_j \in \mathcal{A}_j$ for each $j \neq i$. We construct two CMGs $\mathcal{M}_1, \mathcal{M}_2$. In both, let $U_{i,t}$ be independent fair signs, taking values -1 and 1 with probability $1/2$ each. Their natural actions are

$$X'_{i,t} = \begin{cases} a, & U_{i,t} = 1, \\ b, & U_{i,t} = -1, \end{cases} \qquad X'_{j,t} = c_j \quad (j \neq i).$$

Use independent uniform state noises $U_{s,t+1}$ to sample $P_t(\cdot \mid S_t, \mathbf{X}_t)$ by inverse-transform sampling. Thus the next-state assignment depends only on $(S_t, \mathbf{X}_t, U_{s,t+1})$. All nonconstant noises are mutually independent across stages and components; all remaining latents are constants.

Fix $\alpha > 0$, to be chosen below. Using $\mathbf{1}\{\cdot\}$ for the indicator, define

$$\psi(x_i, u_i) = u_i\big(\mathbf{1}\{x_i = a\} - \mathbf{1}\{x_i = b\}\big).$$

For every player j, set the reward mechanisms to

$$\begin{aligned} f^{(1)}_{Y_j,t}(s, \mathbf{x}, u_i) &= r_{j,t}(s, \mathbf{x}) - \alpha\psi(x_i, u_i), \\ f^{(2)}_{Y_j,t}(s, \mathbf{x}, u_i) &= r_{j,t}(s, \mathbf{x}) + \alpha\psi(x_i, u_i), \end{aligned} \tag{21}$$

where unused latent arguments are suppressed. The binary latent $U_{i,t}$ affects both player i's natural action and the rewards.

Interventional projection. Under $\mathrm{do}(\mathbf{x})$, $U_{i,t}$ remains a fair sign, so

$$\mathbb{E}[\psi(x_i, U_{i,t})] = \big(\mathbf{1}\{x_i = a\} - \mathbf{1}\{x_i = b\}\big)\mathbb{E}[U_{i,t}] = 0.$$

Therefore both models have expected rewards $r_{j,t}$ and transition probabilities P_t, as required by (20). Every joint interventional Markov policy consequently has the same payoffs as in G, so the chosen equilibrium and its payoff $\boldsymbol{\mu}_{\mathrm{NE}}$ are preserved within that policy class.

Natural payoffs. Let $\boldsymbol{\sigma}^*$ choose a, b equally for player i and c_j for each other player. Let μ^*_j be player j's payoff in G under this policy. For $k = 1, 2$, let $\mu_{k,j}$ be player j's natural-policy payoff in $\mathcal{M}_k$, and write $\boldsymbol{\mu}_k = (\mu_{k,j})_{j=1}^n$.

Under natural play, $X_{i,t} = X'_{i,t}$, so $\psi(X_{i,t}, U_{i,t}) = 1$ at every stage. The state/action distribution is that of G under $\boldsymbol{\sigma}^*$. Hence

$$\mu_{1,j} = \mu^*_j - H\alpha, \qquad \mu_{2,j} = \mu^*_j + H\alpha \quad \text{for every } j. \tag{22}$$

All these payoffs are finite. Choosing

$$\alpha > \frac{1}{H} \max_j |\mu_{\mathrm{NE},j} - \mu^*_j|$$

gives $\mu_{1,j} < \mu_{\mathrm{NE},j} < \mu_{2,j}$ for every player, and therefore $\boldsymbol{\mu}_1 \prec \boldsymbol{\mu}_{\mathrm{NE}} \prec \boldsymbol{\mu}_2$. □

Remark C.1 (Equilibrium support). The construction uses the chosen equilibrium only through its payoff vector, so no restriction on its action supports is needed.

Remark C.2 (Magnitude and interpretation). By (22), both gaps $\mu_{\mathrm{NE},j} - \mu_{1,j}$ and $\mu_{2,j} - \mu_{\mathrm{NE},j}$ grow without bound as $\alpha \to \infty$, for every player. This construction allows unrestricted reward magnitudes.

C.2 Proposition C.3: Separation under Bounded Rewards

Proposition C.3 (Separation under bounded rewards)**.** *For every horizon $H \geq 1$, there is a two-player CMG with rewards in $[0, 1]$ whose interventional Markov game has a unique Nash equilibrium with payoff $\boldsymbol{\mu}_{\mathrm{NE}} = (H/2, H/2)$. Its natural policy is its unique CNE, with payoff $\boldsymbol{\mu}_{\mathrm{nat}} = (9H/10, 9H/10)$. Consequently,*

$$\boldsymbol{\mu}_{\mathrm{NE}} \prec \boldsymbol{\mu}_{\mathrm{nat}}, \qquad \|\boldsymbol{\mu}_{\mathrm{nat}} - \boldsymbol{\mu}_{\mathrm{NE}}\|_1 = \frac{4}{5}H,$$

and every player's payoff at every CNE exceeds every player's payoff at every interventional Nash equilibrium by $\frac{2}{5}H$.

Proof. There is one constant state and two actions, $\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$. Draw the agent latents independently across players and stages, with

$$P(U_{i,t} = 0) = P(U_{i,t} = 1) = \frac{1}{2}, \qquad i \in \{1, 2\}, \quad t \in \{1, \dots, H\}, \tag{23}$$

and set the natural actions to $X'_{i,t} = U_{i,t}$. All other exogenous variables are constants. For action and latent vectors $\mathbf{x} = (x_1, x_2)$ and $\mathbf{u} = (u_1, u_2)$, using $\mathbf{1}$ for the indicator, define

$$y_i(\mathbf{x}, \mathbf{u}) = \frac{x_i + 2\mathbf{1}\{x_i = u_i\} + 2\mathbf{1}\{\mathbf{x} = \mathbf{u}\}}{5}, \qquad i = 1, 2,$$

Write $\mathbf{U}_{a,t} = (U_{1,t}, U_{2,t})$. Player i's reward at stage t is $Y_{i,t} = y_i(\mathbf{X}_t, \mathbf{U}_{a,t})$. The three bonuses are summarized in Table S2. Every reward belongs to $[0, 1]$, since its numerator lies between 0 and $1 + 2 + 2 = 5$.

Table S2: Additive components of player i's stage reward. The contributions of all satisfied conditions are added.

Condition	Reward contribution
Choose action 1: $x_i = 1$	$1/5$
Match one's own intuition: $x_i = u_i$	$2/5$
Both players match their intuitions: $\mathbf{x} = \mathbf{u}$	$2/5$

Interventional equilibrium. Consider one stage and omit the index t. Under a joint intervention $\mathrm{do}(\mathbf{x})$, $P(U_i = x_i) = 1/2$ and $P(\mathbf{U}_a = \mathbf{x}) = 1/4$. Therefore

$$\mathbb{E}[Y_i \mid \mathrm{do}(\mathbf{x})] = \frac{x_i + 2(1/2) + 2(1/4)}{5} = \frac{3}{10} + \frac{x_i}{5}.$$

Action 1 strictly dominates action 0, independently of the opponent's action. Hence the unique interventional equilibrium, including mixed strategies, is $(1, 1)$, with payoff $1/2$ to each player.

Counterfactual equilibrium. Fix player i's intuition u_i, and let $j = 3 - i$ be the other player. For every opponent action x_j and latent value u_j, changing player i's action from $1 - u_i$ to u_i increases its reward by

$$\frac{1 + 2u_i + 2\mathbf{1}\{x_j = u_j\}}{5} \geq \frac{1}{5}.$$

Thus following one's intuition is uniquely optimal at each type against every opponent policy, including randomized policies. Both types have positive probability, so the unique CNE is $X_i = U_i$ for both players. It is also the natural policy, and each player's expected reward is

$$\mathbb{E}[Y_i] = \frac{\mathbb{E}[U_i] + 2 + 2}{5} = \frac{9}{10}.$$

Horizon and gaps. In the paper's Markov policy classes, changing a decision rule at one stage affects only that stage's reward: the state is constant and the latents are drawn afresh. The unique stage equilibria therefore repeat at every stage. Multiplying the stage payoffs by H gives the stated vectors. Their coordinate gap is $(9/10 - 1/2)H = 2H/5$, and their ℓ_1 gap is $4H/5$. □

D Proofs for Section 3.1

We prove Theorem 3.2 for CNash-VI-FO with factored estimation. Throughout, $\mathcal{S}$ is the finite state space, $S = |\mathcal{S}|$, and player $i \in \{1, \dots, n\}$ has action space $\mathcal{A}_i$ of size A_i. Write

$$\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i, \qquad A = \prod_{i=1}^{n} A_i, \qquad \widetilde{\mathcal{S}} = \mathcal{S} \times \mathcal{A}.$$

The horizon is $H \geq 1$, every episode starts at s_1 with fresh exogenous draws independent of previous episodes, and $0 \leq Y_{i,t} \leq 1$ almost surely. A product counterfactual policy has joint action probabilities

$$\boldsymbol{\pi}_t(\mathbf{x} \mid s, \mathbf{x}') = \prod_{i=1}^{n} \pi_{i,t}(x_i \mid s, x'_i).$$

Thus each player uses only the state and its own intuition; Π_i^{CTF} denotes this policy class for player i. The shared latent U_t^C may have any distribution allowed by the CMG definition; no independence between players' intuitions is assumed. If $A = 1$, the unique policy is already an exact CNE; below we assume $A \geq 2$.

D.1 Conditional Model

Definition D.1 (Conditional model and augmented state)**.** Write $\widetilde{s} = (s, \mathbf{x}')$ for an augmented state and $\rho_t(\mathbf{x}' \mid s) = P(\mathbf{X}'_t = \mathbf{x}' \mid S_t = s)$ for the joint intuition distribution. Define

$$\widetilde{r}_{t,i}(\widetilde{s}, \mathbf{x}) = \mathbb{E}[Y_{i,t} \mid S_t = s, \mathbf{X}'_t = \mathbf{x}', do(\mathbf{x})], \tag{24}$$

and

$$P_t(s' \mid \widetilde{s}, \mathbf{x}) = P(S_{t+1} = s' \mid S_t = s, \mathbf{X}'_t = \mathbf{x}', do(\mathbf{x})).$$

Here $do(\mathbf{x})$ fixes the executed actions $\mathbf{X}_t = \mathbf{x}$ and leaves the intuition mechanisms unchanged. At a row with $\rho_t(\mathbf{x}' \mid s) = 0$, choose an arbitrary reward mean in $[0, 1]$ and an arbitrary next-state distribution. For $t < H$, the augmented transition is

$$\widetilde{P}_t((s', \mathbf{x}'') \mid (s, \mathbf{x}'), \mathbf{x}) = P_t(s' \mid (s, \mathbf{x}'), \mathbf{x})\rho_{t+1}(\mathbf{x}'' \mid s'). \tag{25}$$

Given the state and full intuition vector, executed actions are drawn using policy randomness independent of the remaining exogenous uncertainty. Hence, conditioning on those actions yields the same reward mean and transition probabilities as the displayed intervention, regardless of the policy. Fresh next-stage intuition latents give (25), with the full joint distribution ρ_{t+1}. These conditional Markov properties follow from Lemma B.3; they do not require independence between reward and next-state observations.

For a function $F : \mathcal{S} \to \mathbb{R}$, write

$$[P_t F](\widetilde{s}, \mathbf{x}) = \sum_{s'} P_t(s' \mid \widetilde{s}, \mathbf{x})F(s').$$

Use the analogous finite sums for empirical transitions and for functions on the augmented state space. The algorithm estimates P_t and ρ_t separately, so it need not estimate a distribution over the SA augmented successor states.

D.2 Value Functions

For a fixed product policy $\boldsymbol{\pi}$, set $W^{\boldsymbol{\pi}}_{H+1,i} = 0$ and define backwards

$$Q^{\boldsymbol{\pi}}_{t,i}(\widetilde{s}, \mathbf{x}) = \widetilde{r}_{t,i}(\widetilde{s}, \mathbf{x}) + [P_t W^{\boldsymbol{\pi}}_{t+1,i}](\widetilde{s}, \mathbf{x}), \tag{26}$$

$$V^{\boldsymbol{\pi}}_{t,i}(s, \mathbf{x}') = \sum_{\mathbf{x}} \boldsymbol{\pi}_t(\mathbf{x} \mid s, \mathbf{x}')Q^{\boldsymbol{\pi}}_{t,i}((s, \mathbf{x}'), \mathbf{x}), \tag{27}$$

$$W^{\boldsymbol{\pi}}_{t,i}(s) = \sum_{\mathbf{x}'} \rho_t(\mathbf{x}' \mid s)V^{\boldsymbol{\pi}}_{t,i}(s, \mathbf{x}'). \tag{28}$$

Here W is the expected remaining reward before the current intuitions are drawn; V additionally conditions on their full vector. This conditioning is used for analysis and does not enlarge a player's information. All three true values lie in $[0, H - t + 1]$. By (25), for $t < H$,

$$Q^{\boldsymbol{\pi}}_{t,i} = \widetilde{r}_{t,i} + \widetilde{P}_t V^{\boldsymbol{\pi}}_{t+1,i} = \widetilde{r}_{t,i} + P_t W^{\boldsymbol{\pi}}_{t+1,i}. \tag{29}$$

All continuation terms at $t = H$ are zero; no intuition at $H + 1$ is needed.

D.3 The Bayesian Equilibrium Oracle

Let $\Delta(\mathcal{A}_i)$ be the probability simplex on $\mathcal{A}_i$. At a fixed state s, let $q(\cdot \mid s)$ be a joint type distribution and let F be a payoff table on $\widetilde{\mathcal{S}} \times \mathcal{A}$. For a product rule $\boldsymbol{\eta}$, define

$$\mathbb{E}_{q,\boldsymbol{\eta}}[F] = \sum_{\mathbf{x}',\mathbf{x}} q(\mathbf{x}' \mid s) \prod_j \eta_j(x_j \mid s, x'_j)F((s, \mathbf{x}'), \mathbf{x}).$$

The expectation first draws the joint intuition and then the executed actions.

Definition D.2 (Stage Bayesian game and BNE oracle)**.** Given prior $q(\cdot \mid s)$ and payoff tables F_i, the stage Bayesian game has players i, private types $x'_i \in \mathcal{A}_i$, and actions $x_i \in \mathcal{A}_i$. The oracle $\textsc{BayesNash}(q, \{F_i\})$ returns a product rule $\boldsymbol{\pi}$ such that, for every player i and every alternative rule $\nu_i : \mathcal{A}_i \to \Delta(\mathcal{A}_i)$,

$$\mathbb{E}_{q,\boldsymbol{\pi}}[F_i] \geq \mathbb{E}_{q,(\nu_i,\boldsymbol{\pi}_{-i})}[F_i]. \tag{30}$$

The state argument of ν_i is fixed here. To express this condition by type, let q_i be the marginal of q. For $q_i(x'_i \mid s) > 0$, define $q(\mathbf{x}'_{-i} \mid s, x'_i) = q(\mathbf{x}' \mid s)/q_i(x'_i \mid s)$ and the interim payoff

$$u_i(x_i; x'_i, \boldsymbol{\pi}_{-i}) = \sum_{\mathbf{x}'_{-i}, \mathbf{x}_{-i}} q(\mathbf{x}'_{-i} \mid s, x'_i) \prod_{j \neq i} \pi_j(x_j \mid s, x'_j) F_i((s, \mathbf{x}'), (x_i, \mathbf{x}_{-i})). \tag{31}$$

Condition (30) is equivalent to best-response optimality at every type of positive probability. Rules at types of probability zero are arbitrary.

Indeed, typewise optimality implies (30) by averaging, and an improvement at a positive-probability type would improve the rule that changes only that type. This formulation uses the joint prior and permits correlated types.

D.4 CNash-VI-FO with Factored Estimation

Fix an episode budget $K \geq 1$ and failure probability $p \in (0, 1)$. Before episode k, let $N_t^k(\widetilde{s}, \mathbf{x})$ count visits to the indicated state–intuition–action row in completed episodes, and let $n_t^k(s)$ count visits to s at step t. Let $\hat{r}_{t,i}^k$ and $\hat{P}_t^k$ be the empirical reward mean and real-state transition at that row. Estimate the joint prior $\hat{\rho}_t^k(\cdot \mid s)$ from all $n_t^k(s)$ observed intuition vectors. At zero counts, use reward estimate zero and arbitrary probability distributions.

Set $c = 16$, write $m \vee 1 = \max\{m, 1\}$, and define

$$\iota = \log \frac{SA^2KH}{p}, \qquad \beta_m = cH\sqrt{\frac{S\iota}{m \vee 1}}, \qquad \gamma_t^k(s) = \min\left\{H, cH\sqrt{\frac{A\iota}{n_t^k(s) \vee 1}}\right\}. \tag{32}$$

Suppress k in the following updates. Set $\overline{W}_{H+1,i} = \underline{W}_{H+1,i} = 0$. At a visited row with $m = N_t(\widetilde{s}, \mathbf{x})$, use

$$\begin{aligned} \overline{Q}_{t,i} &= \min\{\hat{r}_{t,i} + \hat{P}_t \overline{W}_{t+1,i} + \beta_m, H\}, \\ \underline{Q}_{t,i} &= \max\{\hat{r}_{t,i} + \hat{P}_t \underline{W}_{t+1,i} - \beta_m, 0\}. \end{aligned} \tag{33}$$

Row arguments are suppressed; transition terms at $t = H$ are zero. At unvisited rows, set $\overline{Q}_{t,i} = H$ and $\underline{Q}_{t,i} = 0$. At each (t, s), compute an exact BNE for the prior $\hat{\rho}_t(\cdot \mid s)$ and payoff tables $\overline{Q}_{t,i}$, and set

$$\begin{aligned} \overline{W}_{t,i}(s) &= \min\{H, \mathbb{E}_{\hat{\rho}_t, \boldsymbol{\pi}_t}[\overline{Q}_{t,i}] + \gamma_t(s)\}, \\ \underline{W}_{t,i}(s) &= \max\{0, \mathbb{E}_{\hat{\rho}_t, \boldsymbol{\pi}_t}[\underline{Q}_{t,i}] - \gamma_t(s)\}. \end{aligned} \tag{34}$$

Thus uncertainty in the real-state transition is included in the Q-backup and uncertainty in the intuition distribution in the W-backup. Direct backward induction gives $0 \leq \underline{Q} \leq \overline{Q} \leq H$ and $0 \leq \underline{W} \leq \overline{W} \leq H$. Define the certificate for episode k by

$$\Delta_k = \max_i \big(\overline{W}_{1,i}^k(s_1) - \underline{W}_{1,i}^k(s_1)\big). \tag{35}$$

Algorithm 2 CNash-VI-FO with factored estimation

Require: Horizon H, episodes K, failure probability p, initial state s_1.
Initialize counts to zero, empirical estimates as above, and $\Delta \leftarrow +\infty$.
for episode $k = 1, \dots, K$ **do**
  Set $\overline{W}_{H+1,i} = \underline{W}_{H+1,i} = 0$ for every player.
  for $t = H, H-1, \dots, 1$ **do**
    Compute the bonuses (32) and all Q bounds (33).
    for each $s \in \mathcal{S}$ **do**
      $\boldsymbol{\pi}_t(\cdot \mid s, \cdot) \leftarrow \textsc{BayesNash}(\hat{\rho}_t(\cdot \mid s), \{\overline{Q}_{t,i}((s,\cdot),\cdot)\}_i)$.
      Compute $\overline{W}_{t,i}(s), \underline{W}_{t,i}(s)$ using (34).
    end for
  end for
  Compute Δ_k using (35).
  if $\Delta_k < \Delta$ **then**
    $\Delta \leftarrow \Delta_k$; $\boldsymbol{\pi}^{\mathrm{out}} \leftarrow \boldsymbol{\pi}$.
  end if
  Start an episode at s_1.
  for $t = 1, \dots, H$ **do**
    Each player i observes $(s_t, x'_{i,t})$ and draws $x_{i,t} \sim \pi_{i,t}(\cdot \mid s_t, x'_{i,t})$.
    After the actions, log $(s_t, \mathbf{x}'_t, \mathbf{x}_t, \mathbf{y}_t, s_{t+1})$.
  end for
  Update $N_t, n_t, \hat{r}_{t,i}, \hat{P}_t, \hat{\rho}_t$ from the episode log.
end for
return $\boldsymbol{\pi}^{\mathrm{out}}$.

Each logged tuple directly supplies a real-state transition observation. The algorithm never requires another player's intuition before acting, or a next-stage intuition to update a transition.

D.5 The Confidence Gap Bounds Exploitability

Let $I = \{i : A_i \geq 2\}$ be the players who can change their actions. It is nonempty when $A \geq 2$, and $|I| \leq \log_2 A$. Let $\mathcal{E}$ be the event that, simultaneously over episodes, steps and all indicated rows or states,

$$|\hat{r}^k_{t,i} - \widetilde{r}_{t,i}| + H\|\hat{P}^k_t - P_t\|_1 \leq \beta_{N^k_t} \qquad (i \in I,\ N^k_t > 0), \tag{36}$$

$$\sup_{f:\mathcal{A}\to[0,H]} \left| \sum_{\mathbf{x}'} (\hat{\rho}^k_t - \rho_t)(\mathbf{x}' \mid s) f(\mathbf{x}') \right| \leq \gamma^k_t(s). \tag{37}$$

The first line suppresses row arguments and omits the transition term at $t = H$. Since I is nonempty, it also bounds $H\|\hat{P}^k_t - P_t\|_1$ independently of the player. The next subsection proves $P(\mathcal{E}) \geq 1 - p/2$.

For a fixed episode policy $\boldsymbol{\pi}^k$, define player i's best-response value against the other players' fixed policies by

$$B^k_{t,i}(s) = \sup_{\pi'_i \in \Pi^{\mathrm{CTF}}_i} W^{(\pi'_i, \boldsymbol{\pi}^k_{-i})}_{t,i}(s), \qquad B^k_{H+1,i} = 0.$$

Here the supremum concerns the player's remaining decision rules from stage t onward.

Proposition D.3 (Confidence-gap certificate)**.** *On $\mathcal{E}$, every episode k satisfies*

$$\underline{W}^k_{t,i}(s) \leq W^{\boldsymbol{\pi}^k}_{t,i}(s) \leq B^k_{t,i}(s) \leq \overline{W}^k_{t,i}(s) \qquad (i \in I).$$

Consequently, for all players,

$$\max_i \sup_{\pi'_i \in \Pi^{\mathrm{CTF}}_i} \left(W^{(\pi'_i, \boldsymbol{\pi}^k_{-i})}_{1,i}(s_1) - W^{\boldsymbol{\pi}^k}_{1,i}(s_1) \right) \leq \Delta_k.$$

Proof. Fix k and $i \in I$, and suppress k. Fresh intuitions and finite action sets give the best-response recursion

$$B_{t,i}(s) = \max_{\nu_i} \mathbb{E}_{\rho_t, (\nu_i, \boldsymbol{\pi}_{-i,t})}[\widetilde{r}_{t,i} + P_t B_{t+1,i}],$$

where ν_i ranges over current rules that use only s and x'_i. This follows by choosing an optimal continuation at every successor state and then an optimal action at each current type.

Induct backwards from the zero terminal values. If the asserted bounds hold at $t+1$, the model confidence bound and (33) give

$$\overline{Q}_{t,i} \geq \widetilde{r}_{t,i} + P_t B_{t+1,i}, \qquad \underline{Q}_{t,i} \leq Q^{\boldsymbol{\pi}}_{t,i}.$$

Clipping and unvisited-row initialization preserve these bounds because true remaining rewards lie in $[0, H]$. The prior confidence bound and empirical BNE condition now give

$$\begin{aligned} B_{t,i}(s) &\leq \max_{\nu_i} \mathbb{E}_{\rho_t,(\nu_i,\boldsymbol{\pi}_{-i,t})}[\overline{Q}_{t,i}] \\ &\leq \max_{\nu_i} \mathbb{E}_{\hat{\rho}_t,(\nu_i,\boldsymbol{\pi}_{-i,t})}[\overline{Q}_{t,i}] + \gamma_t(s) \\ &= \mathbb{E}_{\hat{\rho}_t,\boldsymbol{\pi}_t}[\overline{Q}_{t,i}] + \gamma_t(s). \end{aligned}$$

Since $B_{t,i} \leq H$, the upper clipping in (34) preserves this inequality. Similarly,

$$\mathbb{E}_{\hat{\rho}_t,\boldsymbol{\pi}_t}[\underline{Q}_{t,i}] - \gamma_t(s) \leq \mathbb{E}_{\rho_t,\boldsymbol{\pi}_t}[\underline{Q}_{t,i}] \leq W^{\boldsymbol{\pi}}_{t,i}(s),$$

and clipping at zero preserves the lower bound. This proves the induction. Subtracting the initial bounds proves the certificate for $i \in I$; players outside I have zero unilateral improvement. □

D.6 Proof of Theorem 3.2

Proof of Theorem 3.2. The case $A = 1$ is immediate. For $A \geq 2$, we establish the confidence event and then bound the smallest certificate.

Confidence bounds. At successive visits to a row, rewards have conditional mean $\widetilde{r}_{t,i}$ and next states have conditional distribution P_t, as shown in Appendix D.1. At each visit to (t, s), the conditional distribution of the full intuition vector, given the preceding observations, is $\rho_t(\cdot \mid s)$. Martingale Hoeffding bounds therefore apply at successive visits, including when the final visit counts are random. For a categorical distribution, apply these bounds to indicators of all subsets of outcomes to obtain an ℓ_1 bound.

There are HSA^2 transition rows, HS intuition rows, and at most K visit prefixes for each. Reward bounds are needed only for $i \in I$, with $|I| \leq \log_2 A \leq A$. A union bound, with absolute constants absorbed into the confidence radii, gives simultaneously at all positive counts

$$|\hat{r}^k_{t,i} - \widetilde{r}_{t,i}| \leq 2\sqrt{\frac{\iota}{N^k_t}}, \qquad \|\hat{P}^k_t - P_t\|_1 \leq 4\sqrt{\frac{S\iota}{N^k_t}}, \qquad \|\hat{\rho}^k_t - \rho_t\|_1 \leq 4\sqrt{\frac{A\iota}{n^k_t}},$$

with probability at least $1 - p/2$; the first bound is for $i \in I$, and row or state arguments are suppressed. The extra union over I costs only a constant factor in $\iota = \log(SA^2KH/p)$. For distributions $q, \hat{q}$ and $f \in [0, H]$,

$$\left| \sum_z (\hat{q} - q)(z) f(z) \right| \leq \frac{H}{2} \|\hat{q} - q\|_1.$$

Since $c = 16$, these inequalities give (36)–(37). At zero state counts, $\gamma^k_t(s) = H$ makes the prior bound automatic. This proves $P(\mathcal{E}) \geq 1 - p/2$. The bounds are uniform over bounded functions, so they also apply to the data-dependent policies and values used by the algorithm.

An episode's certificate. On $\mathcal{E}$, define for every player

$$g^k_{t,i} = \overline{W}^k_{t,i} - \underline{W}^k_{t,i} \in [0, H], \qquad b^k_t(\widetilde{s}, \mathbf{x}) = \min\{H, \beta_{N^k_t(\widetilde{s},\mathbf{x})}\}.$$

At a visited row with $\beta_{N^k_t} < H$, the backups and transition confidence give

$$\overline{Q}^k_{t,i} - \underline{Q}^k_{t,i} \leq \hat{P}^k_t g^k_{t+1,i} + 2\beta_{N^k_t} \leq P_t g^k_{t+1,i} + 3b^k_t.$$

At every other row, including unvisited rows, the final inequality follows from $\overline{Q} - \underline{Q} \leq H$. Applying the prior bound to the action-averaged Q width in (34) gives

$$g^k_{t,i}(s) \leq \mathbb{E}_{\rho_t,\boldsymbol{\pi}^k_t}[P_t g^k_{t+1,i} + 3b^k_t] + 3\gamma^k_t(s). \tag{38}$$

Reward estimates cancel from this width calculation, so it holds for players outside I as well.

Along the actual trajectory of episode k, write $\widetilde{S}_t = (S_t, \mathbf{X}'_t)$ and set

$$Z_k = \sum_{t=1}^{H} [b_t^k(\widetilde{S}_t, \mathbf{X}_t) + \gamma_t^k(S_t)].$$

Let $\mathbb{E}_k$ denote conditional expectation given all information available before executing episode k, including its selected policy. Recursing (38) gives $g_{1,i}^k(s_1) \le 3\mathbb{E}_k Z_k$ for every player. Hence $\Delta_k \le 3\mathbb{E}_k Z_k$ without a multiplicative factor of n.

Summing over episodes. Since $0 \le Z_k \le 2H^2$, the differences $\mathbb{E}_k Z_k - Z_k$ form a bounded martingale difference sequence. Azuma's inequality gives, with probability at least $1 - p/2$,

$$\sum_{k=1}^{K} \mathbb{E}_k Z_k \le \sum_{k=1}^{K} Z_k + 2H^2\sqrt{2K\log(2/p)}.$$

This probability statement is applied before intersecting with $\mathcal{E}$. For a row visited M times, its preceding counts are $0, \dots, M-1$ and $\sum_{m=0}^{M-1}(m \vee 1)^{-1/2} \le 2\sqrt{M}$. Let N_t^{K+1} and n_t^{K+1} be the final counts. At each step their sums are K, so Cauchy–Schwarz gives

$$\sum_{k,t} b_t^k(\widetilde{S}_t, \mathbf{X}_t) \le 2cH\sqrt{S\iota} \sum_{t,\widetilde{s},\mathbf{x}} \sqrt{N_t^{K+1}(\widetilde{s}, \mathbf{x})} \le 2cH^2 SA\sqrt{K\iota},$$

$$\sum_{k,t} \gamma_t^k(S_t) \le 2cH\sqrt{A\iota} \sum_{t,s} \sqrt{n_t^{K+1}(s)} \le 2cH^2\sqrt{SAK\iota}.$$

Intersecting the two probability events and using $\log(2/p) \le \iota$ and $\sqrt{SA} \le SA$, we obtain, for an absolute constant C_0,

$$\min_{1 \le k \le K} \Delta_k \le \frac{1}{K}\sum_{k=1}^{K} \Delta_k \le C_0 H^2 SA\sqrt{\frac{\iota}{K}}$$

with probability at least $1 - p$. The algorithm returns a policy attaining the smallest certificate. Proposition D.3 therefore proves the theorem whenever

$$K \ge C\frac{H^4 S^2 A^2 \iota}{\epsilon^2}, \tag{39}$$

where $C = C_0^2$. This proof requires neither uniform visitation nor a positive lower bound on intuition probabilities. □

D.7 Computational Cost of the Bayesian Equilibrium Oracle

Theorem 3.2 bounds the number of sampled episodes and assumes an exact BNE oracle for the empirical stage Bayesian game. On $\mathcal{E}$, a certified approximate oracle can also be used: for $\eta \ge 0$, if every positive-probability empirical type can improve its interim payoff by at most η, then the certificate for episode k becomes

$$\Delta_k^{(\eta)} := \max_i \big(\overline{W}_{1,i}^k(s_1) - \underline{W}_{1,i}^k(s_1)\big) + H\eta.$$

Indeed, the same backward induction gives $B_{t,i}^k \le \overline{W}_{t,i}^k + (H - t + 1)\eta$ for $i \in I$, while the lower bound is unchanged. Thus an unadjusted certificate of at most ϵ gives an $(\epsilon + H\eta)$-CNE. No-regret dynamics alone do not certify a BNE; their correlated output has the alternative guarantee stated below.

Size and best responses. A pure strategy in a stage Bayesian game is a map $a_i : \mathcal{A}_i \to \mathcal{A}_i$. Its strategic normal form therefore has $A_i^{A_i}$ strategies for player i and $\prod_i A_i^{A_i}$ pure profiles. Planning makes SH oracle calls per episode. Against fixed pure opponent maps, a best response can be computed separately at each own type, in $O(A_i^2 \prod_{j \ne i} A_j)$ operations. To enumerate pure BNEs in the two-player case, enumerate player 1's maps, compute player 2's best-response maps, and

retain the profiles in which player 1 is also best-responding. Unique interim best responses give one candidate per enumerated map; with ties, all relevant best-response maps must be considered. If no pure BNE exists, a mixed equilibrium of the strategic normal form must be computed. Against mixed opponent policies, directly computing all interim action payoffs costs $O(A^2)$ per player, where $A = \prod_i A_i$.

A correlated alternative. A correlated counterfactual policy draws, at each state and step, a fresh joint profile of maps $a_i : \mathcal{A}_i \to \mathcal{A}_i$, using fresh policy randomness independent of the environment's exogenous variables. Player i executes its assigned map at its own intuition. Its Q, V, W values use the same Bellman recursions as above, with the conditional joint action probability $\sigma_t(\mathbf{x} \mid s, \mathbf{x}')$ replacing the product rule. Write $\boldsymbol{\sigma}$ for this policy and $\boldsymbol{\sigma}_{-i}$ for the marginal distribution of the other players' maps. It is an ϵ-Causal CCE (causal coarse correlated equilibrium) if

$$W_{1,i}^{(\pi_i', \boldsymbol{\sigma}_{-i})}(s_1) - W_{1,i}^{\boldsymbol{\sigma}}(s_1) \le \epsilon \qquad \text{for every } i \text{ and } \pi_i' \in \Pi_i^{\mathrm{CTF}}.$$

Here the deviating policy ignores the device's recommendation.

Remark D.4 (CNash-VI-FO with a Bayes-CCE oracle). At every state and step, replace the BNE oracle by an η-Bayes-CCE of the empirical stage game with payoffs $\overline{Q}_{t,i} \in [0, H]$, where the error is measured against unilateral complete-map deviations. Use its correlated joint policy in both the value updates and execution. With the confidence bonuses in (32) and the certificate $\Delta_k^{(\eta)}$ above, an unadjusted certificate of at most ϵ gives an $(\epsilon + H\eta)$-Causal CCE. Under the episode bound of Theorem 3.2, this holds with probability at least $1 - p$.

For $\eta > 0$, such a stage oracle can be implemented by running one full-information Hedge learner for each player and each positive-probability empirical type. Let $p_i^\ell(\cdot \mid x_i')$ be its action distribution at round ℓ. After T rounds, return

$$\sigma_t(\mathbf{x} \mid s, \mathbf{x}') = \frac{1}{T} \sum_{\ell=1}^{T} \prod_i p_i^\ell(x_i \mid x_i').$$

Each round costs $O(nA^2)$ operations when the interim payoffs are computed exactly against the current mixed opponent policies, and $T = \lceil 2H^2 \log(\max_i A_i)/\eta^2 \rceil$ rounds suffice when $\max_i A_i \ge 2$. If every $A_i = 1$, the unique profile is already an exact equilibrium. Execution samples a common index ℓ uniformly using independent policy randomness and gives each player its corresponding rule. The index is sampled afresh at each stage; players do not independently sample the marginal time averages. To implement these stochastic rules as the maps defined above, independently sample each player's map entries according to p_i^ℓ before the intuitions are drawn. Sampling private action coins at the realized type gives the same outcome distribution.

Proof. Fix a state and step. For a positive-probability empirical type x_i', let $g_{i,x_i'}^\ell(x_i) \in [0, H]$ be the interim payoff of action x_i at round ℓ, computed using the empirical conditional type distribution and the other players' current policies. Hedge with a suitable learning rate satisfies

$$\max_{x_i} \sum_{\ell=1}^{T} g_{i,x_i'}^\ell(x_i) - \sum_{\ell=1}^{T} \sum_{x_i} p_i^\ell(x_i \mid x_i') g_{i,x_i'}^\ell(x_i) \le H\sqrt{2T \log A_i}.$$

Let $\hat{\rho}_{i,t}$ be player i's marginal of the empirical joint prior $\hat{\rho}_t$. Weight the inequality by $\hat{\rho}_{i,t}(x_i' \mid s)$ and sum over x_i'. A complete-map deviation chooses one action separately at each type, so its gain against the returned mixture is at most $H\sqrt{2 \log A_i / T}$. This proves the stage guarantee and the stated choice of T.

For the sequential guarantee, the empirical stage inequality used in the backward-induction proof now has an additive η. The same confidence bounds apply to averages under a correlated joint policy, and the maps are drawn using independent policy randomness. At most H such stage errors accumulate. Consequently every unilateral counterfactual deviation gains at most $\max_i(\overline{W}_{1,i}^k(s_1) - \underline{W}_{1,i}^k(s_1)) + H\eta$. The certificate bound proved above completes the proof. □

D.8 Role of Post-hoc Observability

Post-hoc observation of the full intuition vector allows the learner to estimate rewards and transitions conditional on $(S_t, \mathbf{X}_t', \mathbf{X}_t)$, as required by the algorithm above. When opponents' intuitions are not

observed, these conditional tables cannot in general be estimated from the same visitation counts. Section 3.2 studies an alternative identification approach under additional structural assumptions.

E Proofs for Section 3.2

We prove Theorem 3.3 for two players, allowing the shared confounder U_t^C. We use the conditional model and value functions of Appendix D.1–D.2, with rewards in $[0, 1]$ and independent exploration episodes starting at s_1. During exploration, both players choose executed actions uniformly, using independent policy randomness. Agent 1's log contains its own intuition, both executed actions, both rewards, and the next state, as in Algorithm 1. Transition assumptions and estimates are needed only for $t < H$.

E.1 Identifiability Conditions

Let P^{unif} denote probability under uniform exploration and let $\mathcal{R}_t = \{s : P^{\text{unif}}(S_t = s) > 0\}$. Uniform exploration assigns positive probability to every executed action, so $\mathcal{R}_t$ contains every state reachable under a counterfactual policy. Write $\rho_{1,t}$ for the first marginal of the joint intuition distribution ρ_t. For $s \in \mathcal{R}_t$ and $\rho_{1,t}(x_1' \mid s) > 0$, define

$$\rho_{2,t}(j \mid s, x_1') = P(X_{2,t}' = j \mid S_t = s, X_{1,t}' = x_1'), \qquad j \in \mathcal{A}_2.$$

Fix such a state, step, and own intuition. An observed cell is $c = (s, x_1', x_1, x_2)$, with $\mathbf{x} = (x_1, x_2)$. Suppressing (c, t), write

$$w_j = \rho_{2,t}(j \mid s, x_1'), \qquad p_j = P(\mathbf{Y}_t \in \cdot \mid c, X_{2,t}' = j), \qquad P_j = P_t(\cdot \mid (s, x_1', j), \mathbf{x}),$$

where $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})$ is the observed reward pair and conditioning on c abbreviates conditioning on the corresponding observations during exploration. Only j with $w_j > 0$ matter. Uniform action sampling and the conditional model give

$$P(\mathbf{Y}_t \in \cdot \mid c) = \sum_j w_j p_j, \qquad P(S_{t+1} \in \cdot \mid c) = \sum_j w_j P_j.$$

The weights are constant across (x_1, x_2) at fixed (s, t, x_1'); they need not be constant across x_1'. The two conditional reward means are $\widetilde{r}_{t,i}((s, x_1', j), \mathbf{x}) = \int y_i \, p_j(d\mathbf{y})$, for $i = 1, 2$.

Assumption E.1 (Reward family)**.** There is a known family $\mathcal{F}$ of probability distributions on $[0, 1]^2$ containing every component p_j. Any collection of at most $2A_2$ distinct members of $\mathcal{F}$ is linearly independent: a relation $\sum_k b_k f_k = 0$ between these distributions, viewed as signed measures, implies $b_k = 0$ for every k.

Assumption E.2 (Distinctness)**.** Within each reachable (s, t, x_1') with positive own-type probability, the positive weights w_j are pairwise distinct. Each cell is either informative, meaning that it has at least two positive-weight components and their reward-pair distributions p_j are pairwise distinct, or degenerate, meaning that both p_j and, for $t < H$, P_j are constant over the positive-weight types.

Assumption E.3 (Consistently labeled identification)**.** Using the exploration data, including joint observations of $(\mathbf{Y}_t, S_{t+1})$, and any supplied identification information, IDENT_r and IDENT_P consistently estimate the conditional weights $\rho_{2,t}(\cdot \mid s, x_1')$, both conditional reward means $\widetilde{r}_{t,i}$, and, for $t < H$, the real-state transitions P_t on every reachable positive-probability row. Convergence is almost sure, in absolute error for means and in ℓ_1 for distributions. The recovered labels identify the same opponent type across all executed-action cells and all own-intuition values. Any relabeling may depend on the player, state, and step, but not on the other player's intuition, and its conversion must be available when executing the policy.

Transition recovery and common type labels in Assumption E.3 are additional requirements, not consequences of the reward-family and distinctness assumptions. Distinct weights align components across action cells within a fixed own-intuition slice; alignment across slices requires further information. With several positive-weight types, a degenerate cell's reward marginal does not identify their separate weights; joint observations or information from other cells may do so. Assumption E.3 expressly requires recovery of these weights.

Lemma E.4 (Reward-mixture identification)**.** *Under Assumptions E.1–E.2, the reward-pair distribution at an informative cell determines its positive-weight pairs (w_j, p_j) up to permutation. With the common labels in Assumption E.3, it therefore determines both conditional reward means. At a degenerate cell, the common reward means and, for $t < H$, the common transition distribution are directly determined by the observed marginals.*

Proof. Subtract two representations of the reward-pair mixture. The result is a vanishing linear combination of at most $2A_2$ distinct members of $\mathcal{F}$, so every coefficient is zero. Thus the distinct component distributions and their total weights agree. An informative mixture has at least two such distributions, so every admissible competing representation is also informative and has distinct components. Integrating each identified p_j gives both reward means. At a degenerate cell, the observed reward and next-state marginals equal their common component distributions, with the transition assertion needed only for $t < H$. □

Joint observations and transition recovery. For $t < H$, let $q_j = P((\mathbf{Y}_t, S_{t+1}) \in \cdot \mid c, X'_{2,t} = j)$ be the joint component distribution on $[0,1]^2 \times \mathcal{S}$. Then

$$P((\mathbf{Y}_t, S_{t+1}) \in \cdot \mid c) = \sum_j w_j q_j, \qquad P_j(s') = q_j([0,1]^2 \times \{s'\}).$$

The component q_j need not factor into its reward and transition marginals. Identification of the q_j therefore supplies transition recovery by marginalization. For example, at an informative cell this identification holds if the candidate joint components belong to a known family whose collections of at most $2A_2$ distinct members are linearly independent: the same subtraction argument identifies (w_j, q_j) up to permutation. This is one sufficient route to identification; consistency, common labels, and weight recovery remain explicit requirements of Assumption E.3.

E.2 Proof of Theorem 3.3

In the identification phase of Algorithm 1, apply the reward-pair identification procedure separately to each observed cell, align the labels, and apply $\textsc{Ident}_P$ to joint reward–next-state observations under Assumption E.3. Express the recovered tables in actual intuition labels using the available conversion. Estimate $\hat{\rho}^K_{1,t}$ from observed own intuitions and obtain $\hat{\rho}^K_{2,t}(\cdot \mid s, x'_1)$ under the same assumption. Form the joint prior

$$\hat{\rho}^K_t(x'_1, j \mid s) = \hat{\rho}^K_{1,t}(x'_1 \mid s)\hat{\rho}^K_{2,t}(j \mid s, x'_1).$$

Use reward estimates in $[0,1]$ and probability estimates in their respective simplexes, with arbitrary valid defaults at unobserved rows. The planning phase computes an exact equilibrium of this estimated model by backward induction using the joint-prior Bayesian games of Definition D.2.

Proof. Every reachable observed cell has positive probability under uniform exploration and hence is sampled infinitely often almost surely. Assumption E.3 gives consistent labeled reward estimates $\hat{r}^K_{t,i}$ and transition estimates $\hat{P}^K_t$. Empirical own-type frequencies and the conditional-weight estimates give $\|\hat{\rho}^K_t - \rho_t\|_1 \to 0$ almost surely at every reachable state. Write $\widehat{\mathcal{M}}_K$ for the resulting model and $\mathcal{M}$ for the true conditional model. In the following maximum, reward and transition errors are evaluated at $((s, \mathbf{x}'), \mathbf{x})$ and the prior error at s. The maximum ranges over $i \in \{1,2\}$, $t \in [H]$, $s \in \mathcal{R}_t$, $\mathbf{x}'$ with $\rho_t(\mathbf{x}' \mid s) > 0$, and $\mathbf{x} \in \mathcal{A}$; omit the transition term at $t = H$. Define

$$e_K = \max_{t,i,s,\mathbf{x}',\mathbf{x}} \left(|\hat{r}^K_{t,i} - \widetilde{r}_{t,i}| + H\|\hat{P}^K_t - P_t\|_1 + H\|\hat{\rho}^K_t - \rho_t\|_1 \right). \tag{40}$$

Finiteness and consistency imply $e_K \to 0$ almost surely.

For any fixed product policy, write $\hat{Q}, \hat{W}$ and Q, W for its values in the two models. They lie in $[0, H]$. Their Bellman recursions use linearity of conditional expectation and require no reward–next-state independence. Compare them by writing

$$\hat{P}_t\hat{W}_{t+1,i} - P_tW_{t+1,i} = (\hat{P}_t - P_t)\hat{W}_{t+1,i} + P_t(\hat{W}_{t+1,i} - W_{t+1,i})$$

and splitting the prior average in the same way, with the estimated value in the prior-error term. The remaining average is under the true ρ_t and P_t, so it involves only positive-probability types and

reachable successor states. Consequently, if d_t is the largest value error over all product policies, players, and $s \in \mathcal{R}_t$, then $d_t \le e_K + d_{t+1}$ and $d_{H+1} = 0$. Thus, including the error in the initial intuition distribution,

$$\sup_{\boldsymbol{\pi}, i} \left| W^{\boldsymbol{\pi}}_{1,i}(s_1; \widehat{\mathcal{M}}_K) - W^{\boldsymbol{\pi}}_{1,i}(s_1; \mathcal{M}) \right| \le H e_K. \tag{41}$$

Let $\hat{\boldsymbol{\pi}}_K$ be the computed equilibrium of $\widehat{\mathcal{M}}_K$. For every player i and counterfactual deviation π'_i, its true gain is at most its estimated gain plus $2He_K$:

$$W^{(\pi'_i, \hat{\boldsymbol{\pi}}_{K,-i})}_{1,i}(s_1; \mathcal{M}) - W^{\hat{\boldsymbol{\pi}}_K}_{1,i}(s_1; \mathcal{M}) \le 2He_K,$$

because the estimated gain is nonpositive. Taking the supremum proves almost-sure convergence of exploitability to zero. The policy space is compact, and exploitability is continuous as a maximum over finitely many pure counterfactual deviations. Hence every limit point is a CNE. □

E.3 Discussion of Identifiability

Examples with bounded rewards. Joint point masses on $[0,1]^2$ form a linearly independent family. So do product beta distributions $\mathrm{Beta}(\alpha_1, b_1) \otimes \mathrm{Beta}(\alpha_2, b_2)$ with fixed $b_1, b_2 > 0$ and varying $(\alpha_1, \alpha_2) \in (0, \infty)^2$. Another example is the family of bivariate Gaussians with fixed positive-definite covariance and varying location vectors, truncated and normalized on $[0,1]^2$. These examples establish the reward-family condition; transition recovery and common labels remain requirements of Assumption E.3.

Labels and pooling. At each (s,t), each player must know a single bijection between its observed intuitions and the recovered labels, valid across action cells and the opponent's intuition values. Conditional-weight ranks alone need not provide this bijection. Since reward components may differ across action cells, identify mixtures separately and align labels before combining estimates of weights shared across those cells.

Transitions and exploration. If the model specifies that P_j is identical across positive-weight opponent types within each (s,t,a_1,x_2) cell, next-state frequencies in that cell estimate the common distribution, even when reward and next state are dependent. Algorithm 1 assumes that both players explore uniformly using independent randomization and ignoring their intuitions. Type-dependent exploration can alter mixture weights across action cells or leave cells unobserved.

E.4 A Finite-Sample Rate in the Atomic Case

The qualitative consistency assumption supplies no finite-sample rate. We now give an explicit sufficient specialization. Suppose $A_2 \ge 2$ and both rewards are deterministic given $(S_t, \mathbf{X}'_t, \mathbf{X}_t)$. At each reachable observed cell, either the reward pairs are distinct for all opponent types or both the reward pair and the transition distribution are independent of that type. Every reachable (s,t,x'_1) slice must contain an informative executed-action cell. Assume full joint intuition support at reachable states and a known correspondence from conditional-weight ranks to actual opponent-type labels in every slice; the weights themselves remain unknown. Define

$$\begin{aligned} \gamma &= \min_{\substack{t,\ s \in \mathcal{R}_t,\ x'_1 \\ j \ne \ell}} \left| \rho_{2,t}(j \mid s, x'_1) - \rho_{2,t}(\ell \mid s, x'_1) \right|, \\ d_{\min} &= \min_{t,\ s \in \mathcal{R}_t,\ x'_1,\ j} \frac{P^{\mathrm{unif}}(S_t = s) \rho_t(x'_1, j \mid s)}{A_1 A_2}. \end{aligned} \tag{42}$$

Here $x'_1 \in \mathcal{A}_1$ and $j, \ell \in \mathcal{A}_2$ range over all types. Assume $\gamma, d_{\min} > 0$. The constant $d_{\min}$ lower-bounds the probability that an episode supplies a sample to any fixed reachable cell and hidden type.

At a cell with A_2 distinct observed reward pairs, label them by empirical frequency rank and the known correspondence, breaking ties arbitrarily, and estimate transitions within each label. With one observed pair, assign it to all types and use the pooled cell transition. Otherwise use arbitrary valid estimates. In each own-type slice, use any cell with A_2 observed pairs to estimate the conditional weights, or a valid default if none exists yet. Estimate the own-type marginal from all state visits and form the joint prior as above.

Proposition E.5 (Finite-sample rate for deterministic rewards)**.** *Under the conditions of this subsection, run Algorithm 1 with the preceding estimates and an exact Bayesian equilibrium oracle. There is an absolute constant $C > 0$ such that, for $\delta \in (0,1)$ and $\epsilon > 0$, the output is an ϵ-approximate CNE with probability at least $1-\delta$ whenever*

$$K \geq C\left(\frac{SH^4\iota}{d_{\min}\epsilon^2} + \frac{\iota}{d_{\min}\gamma^2}\right), \qquad \iota = \log\frac{SA_1A_2H}{\delta}. \tag{43}$$

Proof. For each step and observed cell c, let $n_{c,j}$ count its exploration samples generated by true opponent type j, and let $N_c = \sum_j n_{c,j}$. Chernoff's inequality gives $n_{c,j} \geq Kd_{\min}/2$ simultaneously once $Kd_{\min} \geq C\iota$. At every informative cell, all A_2 reward pairs are then observed. Hoeffding's inequality gives

$$|n_{c,j}/N_c - \rho_{2,t}(j \mid s, x_1')| < \gamma/4$$

simultaneously once $Kd_{\min} \geq C\iota/\gamma^2$. Thus every frequency ranking is correct, and the known correspondence labels every pair correctly. Both reward means are recovered exactly, also at degenerate cells.

Put $\eta = C_1\sqrt{S\iota/(Kd_{\min})}$ for a sufficiently large absolute constant C_1. Multinomial concentration applied to the true type groups gives $\|\hat{P}_t - P_t\|_1 \leq \eta$ uniformly; degenerate cells use their pooled samples. Each informative cell has $N_c \geq A_2Kd_{\min}/2$, and each reachable state has at least $A_1Kd_{\min}/2$ visits. Hence both the conditional-weight estimate from any informative cell and the empirical own-type marginal have ℓ_1 error at most $C_1\sqrt{\iota/(Kd_{\min})} \leq \eta$. Their product-with-conditional construction gives $\|\hat{\rho}_t - \rho_t\|_1 \leq 2\eta$. These concentration bounds are established for the true groups; on the correct-labeling event the implemented groups coincide with them. The bounds hold for all candidate cells, so they also cover the selected informative cell. A union bound over at most $HSA_1^2A_2^2$ type–cell pairs and the state estimates makes all the preceding statements hold with probability at least $1-\delta$, after enlarging C, C_1.

On this event, (40) gives $e_K \leq 3H\eta$. The value comparison (41) and equilibrium transfer therefore bound exploitability by $6H^2\eta$. The first term of (43) makes this at most ϵ, while the second ensures the count and labeling bounds, since $\gamma \leq 1$. This proves the proposition. □

Distinct deterministic reward pairs are observed exactly, so no reward-gap parameter is needed. The γ^{-2} term is a sufficient sampling requirement for frequency-rank labeling; no matching lower bound is claimed. The dependence on $d_{\min}$ reflects uniform exploration.

F Experiments

This appendix expands on Sec. 4. Sec. F.1 covers the CNash-VI-FO experiment. For the windy MPE, Sec. F.2 details the environment, the wind machinery, and the configurations compared; Sec. F.3 specifies the deep MARL algorithm (MADDPG with counterfactual inputs); Sec. F.4 reports the multi-seed results on both tasks, including the wind-conditional breakdown. Sec. F.5 covers the ICPD, including additional opponents.

F.1 CNash-VI-FO on Tabular CMGs

Instances. We use three two-player instances.

- **ICPD.** Ex. 1.1 in M_1 with $H = 5$. The state is trivial ($S = 1$), so the augmented state is the pair of intuitions ($\tilde{S} = 4$). Payoffs are those of Fig. 1a, rescaled from $[-14, 0]$ to $[0, 1]$.
- **Random CMG.** $S = 3$, $A_1 = A_2 = 2$, $H = 4$ ($\tilde{S} = 12$). Mean rewards are drawn from $\mathrm{Beta}(1.5, 1.5)$ and transition rows from $\mathrm{Dir}(0.7 \cdot \mathbf{1})$. Each agent's intuition law $\rho_{i,t}(\cdot \mid s)$ is drawn from $\mathrm{Dir}(3 \cdot \mathbf{1})$, independently across agents, and $\rho_t = \rho_{1,t}\,\rho_{2,t}$.
- **Random CMG with correlated intuitions.** The same rewards, transitions and marginals $\rho_{i,t}$, with the intuitions coupled: $\rho_t(x_1', x_2' \mid s) = \rho_{1,t}(x_1' \mid s)\,\rho_{2,t}(x_2' \mid s) + d\,(-1)^{x_1'+x_2'}$, where $d = 0.9\min_{\mathbf{x}'}\rho_{1,t}\rho_{2,t}$. The correlation between the two intuitions, averaged in absolute value over (t, s), is 0.51.

In all three, realized rewards are Bernoulli with the tabulated mean.

Table S3: CNash-VI-FO at $K = 2 \cdot 10^5$ over 5 seeds (mean $\pm$ std). Slope: fit of $\log \Delta_{\text{out}}$ on $\log K$ over 12 values of K; reference slope -0.50. Violations: checkpoints at which exploitability exceeded Δ_{out}, out of 5×12.

instance	$\tilde{S}$	slope	R^2	Δ_{out}	exploitability	violations
ICPD	4	-0.518	0.977	0.0504 ± 0.0002	0	0/60
random	12	-0.582	0.994	0.1871 ± 0.0013	0.0026 ± 0.0005	0/60
random, correlated	12	-0.572	0.995	0.1843 ± 0.0018	0.0032 ± 0.0006	0/60

Implementation. We run an empirical variant of Alg. 2 with $p = 0.1$, ι as in Thm. 3.2, and bonus $\beta_m = 0.05\sqrt{\tilde{S}H^2\iota/m}$, where $\tilde{S} = |\tilde{\mathcal{S}}| = S\prod_i A_i$. The variant estimates the augmented-state transition jointly, uses no separate intuition-prior bonus ($\gamma \equiv 0$), sets $\overline{Q} = H$, $\underline{Q} = 0$ at unvisited rows and a uniform prior at unvisited states, and reports Δ_{out} as the resulting gap $\max_i(\overline{W}_{1,i} - \underline{W}_{1,i})(s_1)$. These bonuses are much smaller than those in (32), so Δ_{out} does not inherit the high-probability certificate of Thm. 3.2. To reach $K = 2 \cdot 10^5$ episodes in seconds, the algorithm replans at every episode up to $k = 200$ and then each time k grows by 5%, and $\boldsymbol{\pi}^{\text{out}}$ is selected among the replanned policies. BAYESNASH is solved exactly in strategic normal form: a pure strategy maps types to actions, so the stage Bayesian game is an $A_1^{A_1} \times A_2^{A_2}$ bimatrix game whose Nash equilibria are its BNEs.

Measurement. Each run is evaluated at 12 values of K, log-spaced from 10^3 to $2 \cdot 10^5$. With 3 instances and 5 seeds, this gives the 180 seed–checkpoint pairs of Sec. 4.1. At each one we record Δ_{out} and the exploitability of $\boldsymbol{\pi}^{\text{out}}$, the left-hand side of (9). Exploitability is computed exactly on the true model: intuitions are redrawn from $\rho_t(\cdot \mid s)$ at every step (Remark B.2), so with $\boldsymbol{\pi}_{-i}$ fixed, agent i's best response solves an MDP on (s, x_i') by backward induction. The slope is the least-squares fit of $\log \Delta_{\text{out}}$, averaged over seeds, on $\log K$.

Results. Table S3 reports the fits. Exploitability never exceeded Δ_{out}, and at $K = 2 \cdot 10^5$ it is about $70\times$ smaller than the certificate on the random CMGs. On the ICPD it is exactly 0 from $K = 51{,}407$ (two seeds) or $58{,}887$ (three seeds) on.

F.2 Environment: Windy Multi-Particle Environment

We extend two PettingZoo MPE tasks with wind, a disturbance that both moves the objects the agents track and blinds the sensors tracking them.

- `simple_spread` (cooperative): 3 agents, 3 targets, $H = 50$. Each agent receives the negative sum of minimum agent-to-target distances minus its collision penalty, as in the original task.
- `simple_adversary` (mixed): 2 good agents, 1 adversary and 2 targets, one of which is the goal; $H = 75$. The good agents are rewarded when the closest good agent is near the goal and the adversary is far from it; the adversary is rewarded for being near the goal.

Wind. When calm, a gust starts with probability p_{start} per step, in a direction drawn uniformly from left, right, up and down, and lasts T_{wind} steps. During a gust every target receives an additive force of magnitude w_{sens} in the gust direction, and each agent's visible observation entries (positions and velocities of itself and the others, relative target positions) are replaced by -1. The wind also pushes the agents through their natural action: an agent that follows its natural action drifts with the gust, while an agent executing its own chosen action moves as commanded. Parameters and empirical wind statistics are in Tbl. S4.

Natural action. During a gust, a counterfactual agent's observation is extended by its natural action $X'_{i,t}$: the greedy discrete action toward its target (the nearest target in `simple_spread`, the goal in `simple_adversary`), computed from the true state. Outside gusts the slot is zero. The slot is one-hot encoded before it enters the actor and the critic. Non-causal agents receive no slot. In `simple_spread` all agents are counterfactual; in `simple_adversary` only the two good agents are.

Table S4: Wind and environment hyperparameters. Wind statistics are empirical, measured over 1000 random-policy episodes.

parameter	spread	adversary
episode length H	50	75
p_{start} (per-step)	0.05	0.05
T_{wind} (gust duration)	6	6
w_{sens} (wind sensitivity)	5.0	5.0
damping	0.25	0.25
mean wind events / episode	2.0	2.9
$P(\geq 3$ wind events)	0.31	0.62
fraction of steps with active wind	24%	15%

Algorithm 3 Causal MADDPG (the deep MARL instantiation evaluated in Sec. 4).

Require: n agents; actors $\{\pi_{\theta_i}\}$; centralized critics $\{Q_{\phi_i}\}$; targets $\{\pi_{\theta'_i}, Q_{\phi'_i}\}$; replay $\mathcal{D}$; Gumbel–Softmax temperature τ_{GS}; Polyak rate τ; train period T_{train}; warmup T_{warm}; discount γ; learning rates η_Q, η_π.

for env step $= 1, 2, \ldots$ **do**
  Observe s_t; each agent i receives $x'_{i,t} \leftarrow f_{X_i}(s_t, U_{i,t}, U_t^C)$ and local view $o_{i,t}$ with $x'_{i,t}$ appended.
  Each agent i samples $x_{i,t} \sim \text{GS}_{\tau_{\text{GS}}}\big(\pi_{\theta_i}(o_{i,t})\big)$ (straight-through one-hot).
  Execute $\mathbf{x}_t$; observe $\{r_{i,t}\}$, s_{t+1}; store $(\mathbf{o}_t, \mathbf{x}_t, \mathbf{r}_t, \mathbf{o}_{t+1}, d_t)$ in $\mathcal{D}$, where d_t indicates a true terminal state (time-limit truncation is not terminal, so $d_t = 0$ in our tasks).
  if step $\bmod T_{\text{train}} = 0$ **and** $|\mathcal{D}| \geq T_{\text{warm}}$ **then**
    Sample minibatch from $\mathcal{D}$.
    for each agent i **do**
      $y_i \leftarrow r_i + \gamma\,(1-d)\,Q_{\phi'_i}\big(\mathbf{o}',\, \bar{x}'_1, \ldots, \bar{x}'_n\big)$, with $\bar{x}'_j$ the greedy one-hot action of $\pi_{\theta'_j}(o'_j)$ *// target*
      $\phi_i \leftarrow \phi_i - \eta_Q\, \nabla_{\phi_i}\big(Q_{\phi_i}(\mathbf{o}, \mathbf{x}) - y_i\big)^2$ *// critic step*
      $\theta_i \leftarrow \theta_i + \eta_\pi\, \nabla_{\theta_i}\, Q_{\phi_i}\big(\mathbf{o}, x_1, \ldots, \pi_i^{\text{ST}}(o_i), \ldots, x_n\big)$, with $\pi_i^{\text{ST}}(o_i)$ the straight-through one-hot sample *// DPG step*
    end for
    Polyak: $\phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i$, $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$ for all i.
  end if
end for

F.3 Algorithm: MADDPG with Counterfactual Inputs

We use MADDPG [Lowe et al., 2017] with the Causal MADDPG inputs of Sec. 3.3 (Alg. 3). Each actor sees its local observation and, if the agent is counterfactual, its own natural action. Each centralized critic sees $\mathbf{o}_t = (o_{1,t}, \ldots, o_{n,t})$, the concatenated local observations including any natural-action slots (so $\mathbf{x}'_t$ enters through them), and the joint action. All networks are 2×64 ReLU MLPs. Discrete actions use the straight-through Gumbel–Softmax estimator ($\tau_{\text{GS}} = 1$), and target actors act greedily, outputting the one-hot greedy action. After 10^4 warm-up steps, each agent takes one gradient step every 100 environment steps on a minibatch of 1024 transitions from a shared replay buffer, and target networks are updated by Polyak averaging. Time-limit truncations are not treated as terminal: masking the bootstrap there biased Q low and prevented learning. Hyperparameters are in Tbl. S5.

F.4 Full Results

Each configuration is trained with 5 seeds $(2, 3, 4, 5, 42)$ for 50,000 episodes. Every run is evaluated on the same 300 deterministic episodes (evaluation seeds $999{,}000 + i$), so wind realizations are identical across runs, and within-bin comparisons are paired. We report the mean $\pm$ std across seeds of each run's 300-episode mean, and test the gap with a paired t-test over seeds.

Table S5: MADDPG training hyperparameters (shared across environments and configurations).

hyperparameter	value
optimizer	Adam (default β)
actor lr / critic lr	$7\text{e}{-4}$ / $7\text{e}{-4}$
discount γ	0.95
Polyak τ	0.005
replay buffer capacity	1,000,000
batch size	1024
warmup env steps	10,000
`train_every` (env steps)	100
gradient steps per train	1
max grad norm	0.5
Gumbel–Softmax temperature τ_{GS}	1.0
hidden layers $\times$ dim	2×64
activation	ReLU
parallel envs n_{envs}	8
total episodes per run	50,000

Table S6: Aggregate evaluation reward per agent, mean $\pm$ std across 5 seeds. Gap: counterfactual $-$ non-causal, paired over seeds; p: paired t-test over seeds. In `simple_adversary` only the good agents receive the natural action.

task	agents	counterfactual	non-causal	gap	seeds with gap > 0	p
spread	all	-28.64 ± 2.93	-32.28 ± 1.18	$+3.64 \pm 3.03$	$5/5$	0.055
adversary	good	$+9.98 \pm 9.96$	$+8.45 \pm 7.16$	$+1.52 \pm 12.34$	$3/5$	0.80
adversary	adversary	-41.96 ± 8.66	-41.56 ± 9.86	-0.40 ± 14.71	$2/5$	0.95

Per-seed aggregate gaps (seeds $2, 3, 4, 5, 42$) are $+0.67$, $+5.65$, $+1.99$, $+1.97$, $+7.93$ on `simple_spread` and $+2.66$, -4.54, $+0.74$, -12.27, $+21.01$ on `simple_adversary`.

Analysis. *(1) `simple_spread`.* Causal MADDPG is better in every seed, by $+3.64 \pm 3.03$ ($p = 0.055$), and on average in every wind bin (Tbl. S7, Fig. 3f). The gap is largest in the two windiest bins but is already $+3.05$ in episodes without wind, where the natural-action slot is constant. Part of the gain therefore reflects how the policy was trained rather than information used at execution, and the increase from zero to ≥ 3 wind events ($+1.2$) is within seed variation.

(2) `simple_adversary`. With 5 seeds the two configurations do not separate: the good agents' gap is $+1.52 \pm 12.34$ ($p = 0.80$), seeds 3 and 5 reverse, and the per-bin gaps are not ordered by wind. The adversary's own reward is unchanged (-41.96 against -41.56).

F.5 Iterated Causal Prisoner's Dilemma

Instance. The ICPD of Ex. 1.1 in M_1 ($R_{i,t} = 1$, $P(U_{i,t} = 0) = 0.6$ independently across rounds and agents), with the payoffs of Fig. 1a. There is no state beyond the round counter.

CNash-VI-NO. *Exploration:* both agents play uniformly random actions for 10^5 rounds. Agent 1 records its own natural action $x'_{1,t}$, the joint action and the reward pair $(Y_{1,t}, Y_{2,t})$, and never observes $x'_{2,t}$; each identification cell (x'_1, x_1, x_2) has expected count 15,000 if $x'_1 = 1$ and 10,000 if $x'_1 = 0$, since $P(X'_1 = 1) = 0.6$. *Identification:* since there are no state transitions, only rewards need identifying. In each cell the reward-pair law is a mixture over $x'_2 \in \{0, 1\}$. Since the ICPD rewards are deterministic and distinct, either of two procedures recovers the components: (i) identify the two observed reward-pair atoms and their frequencies, and label them by frequency rank using the known correspondence ($P(X'_2 = 1) = 0.6$), as in App. E.4; or (ii) solve for the component means from each cellwise mixture under Asm. E.1, with type labels supplied as required by Asm. E.3. Raw rewards are not pooled across cells (App. E.3). *Exploitation:* BAYESNASH on the recovered payoffs, indexed by (x'_1, x'_2), gives a stationary policy $\hat{\pi}_1(\cdot \mid x'_1)$. It is deployed for 10^4 rounds against always-defect and tit-for-tat, which repeats agent 1's last action (Fig. 3d–e).

Table S7: Per-agent reward by number of wind events per episode, mean ± std across 5 seeds; n is the number of episodes per bin in each run. Gap: counterfactual − non-causal.

task	wind events	n	counterfactual	non-causal	gap	seeds with gap > 0
spread	0	29	-24.73 ± 3.32	-27.78 ± 1.24	$+3.05 \pm 2.90$	$4/5$
	1	79	-25.44 ± 2.94	-28.27 ± 0.72	$+2.84 \pm 2.97$	$5/5$
	2	93	-29.39 ± 3.23	-33.28 ± 1.33	$+3.89 \pm 3.06$	$5/5$
	≥ 3	99	-31.62 ± 3.12	-35.85 ± 1.94	$+4.23 \pm 3.87$	$5/5$
adversary (good)	0	9	$+11.66 \pm 3.03$	$+8.39 \pm 1.51$	$+3.27 \pm 3.69$	$4/5$
	1	24	$+7.09 \pm 3.03$	$+6.35 \pm 2.15$	$+0.73 \pm 3.38$	$3/5$
	2	92	$+8.26 \pm 4.94$	$+6.07 \pm 2.91$	$+2.18 \pm 5.42$	$3/5$
	≥ 3	175	$+11.19 \pm 14.10$	$+10.00 \pm 10.58$	$+1.19 \pm 17.85$	$2/5$

Table S8: Long-run payoff per round of $\sigma^\star$ against interventional opponents, exact via the stationary distribution of the chain over last actions; Hedge is evaluated at its limit. Last column: the best payoff agent 1 obtains against the same opponent with a memory-1 interventional policy. Values are identical in M_1 and M_2.

opponent	$\sigma^\star$	opponent	best interventional (agent 1)
always-defect	0.00	-2.40	-1.92
tit-for-tat	-0.10	-3.55	-0.98
win-stay-lose-shift	-0.12	-3.84	-0.98
generous tit-for-tat ($g = 0.3$)	-0.14	-4.07	-0.98
Hedge (no-regret)	0.00	-2.40	-1.92

Uniform dominance. The advantage does not depend on the opponent. Let $\sigma^\star$ be the optimal counterfactual policy ("follow intuition" in M_1), and let agent 2 play any interventional strategy that defects with probability π_{D}. In the stage game, $\sigma^\star$ earns at least 0.24 per round more than every interventional strategy of agent 1. Its margin over agent 1's best interventional response, always-defect, ranges from 0.24 (when agent 2 always cooperates) to 1.92 (when agent 2 always defects). Tbl. S8 extends this to reactive opponents.

The code is available at `https://anonymous.4open.science/r/cmarl26-2026/`.

Advantage over interventional strategies. In the ICPD instance M_1, the latent variables are independent across agents and rounds, with $P(U_{i,t} = 0) = 0.6$, and $X'_{i,t} = 1 - U_{i,t}$. An interventional strategy may use past observations and randomization independent of the current latent variables, but not the current intuitions. Write $v_1(\pi_1, \pi_2)$ for agent 1's expected stage payoff under the policy pair (π_1, π_2).

Proposition F.1 (Uniform advantage of counterfactual play)**.** *Let $\pi_1^\star$ execute $X_{1,t} = X'_{1,t}$. This policy maximizes agent 1's expected stage payoff among policies using its own intuition, against every interventional strategy of agent 2. For every pair of interventional stage strategies σ_1, σ_2,*

$$v_1(\pi_1^\star, \sigma_2) \geq v_1(\sigma_1, \sigma_2) + 0.24.$$

Moreover, for any positive integer T, using $\pi_1^\star$ for T rounds yields at least $0.24T$ more expected cumulative reward than any interventional strategy of agent 1 against the same interventional opponent, even when both interventional strategies depend on the past.

Proof. Suppress the round index. Averaging the ICPD payoffs over U_2, when $U_1 = 0$, action $X_1 = 1$ gives zero reward, whereas $X_1 = 0$ has strictly negative expected reward against either fixed action of agent 2. When $U_1 = 1$, the expected rewards from $X_1 = 0$ and $X_1 = 1$ are respectively -0.6 and -1.2 against $X_2 = 0$, and 0 and -4.8 against $X_2 = 1$. Thus $X_1 = 1 - U_1 = X'_1$ is optimal for either opponent action, and hence for every interventional mixture.

Averaging also over U_1 gives the following exact expected payoffs for agent 1; each column fixes agent 2's executed action independently of the current latent variables:

Agent 1's rule	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	-0.984	-6.96
$X_1 = 1$	-0.48	-1.92
$X_1 = X_1'$	-0.24	0

The last row is at least -0.24 in either column, whereas every entry in the first two rows is at most -0.48. Randomization preserves these bounds, proving the stage-game claim.

For repeated play, let $\mathcal{H}_{t-1}$ denote the history before round t. The current latent variables are independent of this history, so the same table and bounds apply conditionally on $\mathcal{H}_{t-1}$. Consequently, under $\pi_1^\star$ agent 1's conditional expected reward is at least -0.24 in every round, while under any interventional strategy it is at most -0.48. Taking expectations and summing gives the claimed $0.24T$ gap. These bounds hold under each policy pair separately, so the argument permits the opponent to respond differently to their different histories. □

This covers win-stay-lose-shift, generous TfT, and no-regret learners whose actions use only past observations and independent randomization.