TECHNICAL REPORT
R-152
May, 2026

Learning in Causal Markov Games

Aurghya Maiti and **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
{aurghya,eb}@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. Empirically, on a confounded windy variant of the Multi-Particle Environment and the Iterated Causal Prisoner's Dilemma, counterfactual agents strictly dominate their non-causal counterparts.

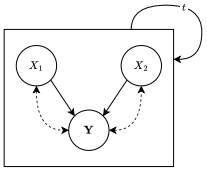
1 Introduction

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. Nash-Q [Hu and Wellman, 2003, Littman et al., 2001], Nash VI and its optimistic variants [Liu et al., 2021], V-learning [Jin et al., 2021], and decentralized policy-gradient methods [Daskalakis et al., 2020] deliver finite-sample regret and PAC guarantees for Nash, correlated, and coarse-correlated equilibria in tabular MGs, with function-approximation extensions [Xie et al., 2020] powering modern MARL at scale.

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

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

(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 Prisoner's Dilemma. (c) Payoffs under different strategies in the Iterated Causal Prisoner's Dilemma

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: two environments that look identical at the interventional level can induce sharply different outcomes once agents reason at the natural or counterfactual layer of the PCH. The example below makes this concrete.

Example 1.1 (Iterated Causal Prisoner's Dilemma)**.** Two friends repeatedly engage in criminal activity. After each offense they are apprehended and questioned separately. In each episode $t \in \{1, \dots, H\}$, 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 episode 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 episodes. Each individual also has an intrinsic ability $R_{i,t} \in \{0, 1\}$ for reading these circumstances, yielding the natural action $X_{i,t} \leftarrow f_X(R_{i,t}, U_{i,t}) = R_{i,t} \oplus U_{i,t}$; when $R_{i,t} = 1$ the agent cooperates iff circumstances are favorable; $R_{i,t} = 0$ reverses this judgment. The latent variables $\{U_{i,t}, R_{i,t}\}_{t=1}^{H}$ and the function f_X are fixed and unknown to the agents. The per-episode 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 episodes in both environments. If individuals ignore their intuition and optimize directly over actions, the induced stage game is the classical Iterated Prisoner's Dilemma in either environment, and the standard Nash equilibrium strategies, such as always defect (D) and tit-for-tat (TfT), coincide [Shoham and Leyton-Brown, 2008].

The two environments diverge sharply once natural intuition is admitted. Mutual natural play yields per-episode 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 (Fig. 1c; derivations in Sec. 2 and Appendix E). Although M_1 and M_2 are indistinguishable at the interventional level, they induce markedly different long-run outcomes once behavioral or counterfactual reasoning is admitted, and in either environment, classical Markov-Game 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 parts of this problem. MDPUC provides such a lift in the single-agent sequential setting, while the Causal Normal-Form Game (CNFG) [Maiti et al., 2025] studies counterfactual policies in single-step multi-agent interactions. Neither captures sequential multi-agent learning under unobserved confounding. 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): the suboptimality of any interventional Nash equilibrium relative to its causal counterpart can be made arbitrarily large.
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-exploit variant for two-player general-sum CMGs that asymptotically converges to an Equilibrium (Sec. 3.2).
3. We provide a counterfactual augmentation for deep MARL (Sec. 3.3) and show empirically that counterfactual agents strictly dominate non-causal baselines 4.

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, $\mathcal{A}_i^{(\cdot)}$ 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.

2 Causal Markov Games

In this section, we introduce *Causal Markov Games*, a framework for sequential multi-agent decision-making in environments that allows for 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. Throughout, we re-use the L_1 (natural) and L_2 (interventional) actions defined in Sec. 1.1 without restating them.

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.

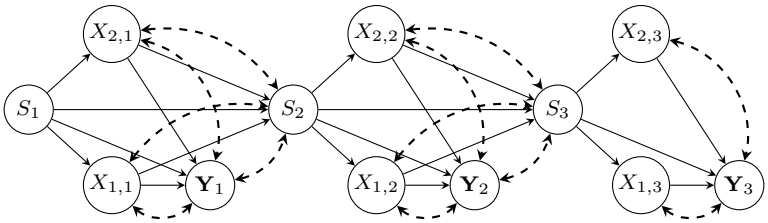


Figure 2: Causal diagram of a Causal Markov Game with $n = 2$ agents over $H = 3$ stages

1. *Exogenous variables.* $\mathbf{U} = \{U_t\}_{t=1}^{H}$, where each $U_t = (\mathbf{U}_{a,t}, U_{s,t}, \mathbf{U}_{y,t})$, $\mathbf{U}_{a,t} = (U_{1,t}, \ldots, U_{n,t})$ and $\mathbf{U}_{y,t} = (U_{y_1,t}, \ldots, U_{y_n,t})$.

2. *Endogenous variables.* $\mathbf{V} = \mathbf{S} \cup \mathbf{X} \cup \mathbf{Y}$, with states $\mathbf{S} = (S_t)_{t=1}^{H}$, 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 $i \in \{1, \ldots, n\}$,
$$\begin{aligned} S_t &\leftarrow f_S(S_{t-1}, \mathbf{X}_{t-1}, \mathbf{U}_{a,t-1}, U_{s,t}), \\ X_{i,t} &\leftarrow f_{X_i}(S_t, U_{i,t}), \\ Y_{i,t} &\leftarrow f_{Y_i}(S_t, \mathbf{X}_t, \mathbf{U}_{a,t}, U_{y_i,t}). \end{aligned} \tag{2}$$

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

The model is initialized by a distribution $P(S_1)$ over the initial state. □

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} depends on $U_{i,t}$, while the reward mechanism f_{Y_i} may read the full vector $\mathbf{U}_{a,t}$, resulting in UC between agents action and the reward; the state S_t depends on previous stage exogenous variables $\mathbf{U}_{a,t-1}$, so there is now unobserved confounding between the agents' action and the next state. Two structural consequences follow. *First*, define the intuition $X'_{i,t} := f_{X_i}(S_t, U_{i,t})$, which is the evaluation of the mechanism in the natural regime. Mutual independence of $\{U_{i,t}\}_{i \in [n]}$ given S_t yields the *conditional independence of intuitions*, namely,

$$X'_{i,t} \perp\!\!\!\perp X'_{j,t} \,\big|\, S_t \qquad \text{for all } i \neq j, \tag{3}$$

which makes counterfactual reasoning naturally decentralized. So, each agent's intuition is an independent signal given the state, so policies that condition on it remain product policies. *Second*, the shared agent latents $\{U_{i,t}\}_{i \in [n]}$ enter 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 rule $\pi_{i,t}$ maps the state S_t and the agent's intuition $X'_{i,t}$ to a distribution over realized actions $X_{i,t} \in \mathcal{A}_i$. A *joint counterfactual policy* is a tuple $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \in \prod_{i=1}^{n} \Pi_i^{\mathrm{CTF}}$.

A joint counterfactual policy is decentralized by construction: agent i's 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.

Now, we demonstrate the modelling and action spaces in CMG through the following example:

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, 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. Per-agent independence of $\{U_{i,t}\}$ holds by construction, so (3) is satisfied.

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-episode payoff is

$$\mathbb{E}[Y_{i,t}] \;=\; \sum_{u_1,u_2} P(u_1)P(u_2)\, f_{Y_i}(u_1, u_2,\, u_1, u_2) \;=\; -2.4, \tag{5}$$

where the sum aggregates the four joint (u_1, u_2) outcomes weighted by the payoff matrix in Fig. 1a. Both agents receive -2.4, giving $\mu_1 = (-2.4, -2.4)$.

L_2 Nash payoff in M_1 and M_2. For any 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

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

The dominant strategy in either environment is always-defect, with Nash payoff $\mu_{\mathrm{NE}} = (-1.9, -1.9)$.

Counterfactual Payoff in M_2. Consider the counterfactual policy $\pi^{\mathrm{opp}}_{i,t}(X_{i,t} = 1 - x' \mid S_t, x') = 1$, 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-episode payoff becomes

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

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 $\mu_1 = (-2.4, -2.4)$. 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 Markov game with $|\mathcal{A}_i| \geq 2$ for some agent i. There exist Causal Markov Games $\mathcal{M}_1$ and $\mathcal{M}_2$, each with a confounder $U_{i,t}$ of support at least 2 on which the natural mechanism f_{X_i} depends non-trivially, that share the same interventional projection G and whose L_1 payoff vectors μ_1, μ_2 satisfy*

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

in the strict Pareto order, where μ_{NE} is the Nash payoff of G.

Moreover, the gap $\|\mu_{\mathrm{NE}} - \mu_1\|_1$ or $\|\mu_2 - \mu_{\mathrm{NE}}\|_1$ can be made arbitrarily large. The proof is provided in Appendix B.1. 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

In Sec. 2, we showed how Causal Markov Games strictly generalizes Markov Games and that the optimal policy may require counterfactual reasoning. In this section, we address the problem of learning from samples, where we develop learning algorithms 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 explore-exploit procedure for two-player games; in Sec. 3.3, we extend deep MARL architectures to the counterfactual action space.

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. First, note that the per-agent latents $\{U_{i,t}\}_{i\in[n]}$ are mutually independent, so the intuitions $X'_{i,t} = f_{X_i}(S_t, U_{i,t})$ are conditionally independent given S_t, which is what makes *product* counterfactual policies the right object to learn. Second, we assume *post-hoc observability*: 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. However, each agent still sees only its own intuition $X'_{i,t}$ before acting, and is blind to the other agents' intuitions.

Under these conditions, the natural actions serve as a proxy for the unobserved exogenous variables: conditioning on $(S_t, \mathbf{X}'_t)$ captures information about U_t relevant to rewards and transitions, given the executed actions. 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 at each step, the randomness in $\boldsymbol{\pi}$, and the state transitions. We write $\boldsymbol{\pi}_{-i} = (\pi_j)_{j\neq i}$ 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:

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

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. CNE is therefore the appropriate solution concept for product counterfactual policies in a CMG. *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 solves the game where each agent's type is its natural action x'_i, drawn from the empirical distribution $\hat{\rho}_{i,t}(\cdot \mid s) \approx P(X'_{i,t} \mid S_t = s)$ estimated from data. 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 selects the policy minimizing the optimism gap $\max_i\big(\overline{W}_{1,i}(s_1) - \underline{W}_{1,i}(s_1)\big)$, where $\overline{W}_{1,i}$ and $\underline{W}_{1,i}$ are optimistic and pessimistic estimates of the expected total reward $W_{1,i}$. The pseudocode and parameter settings are in Appendix C.1.

Theorem 3.2 (Sample Complexity of CNash-VI-FO)**.** *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*

$$K \;\geq\; \Omega\left(\frac{H^4 S^2 \left(\prod_i A_i\right)^3 \iota}{\epsilon^2} \right). \tag{10}$$

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

CNash-VI-FO pays an extra factor of $(\prod_i A_i)^2$ over the $\tilde{O}(H^4 S^2 \prod_i A_i/\epsilon^2)$ bound for standard n-player Markov games [Liu et al., 2021], the cost of folding the natural actions into the state. For two players ($|\mathcal{A}_1| = A$, $|\mathcal{A}_2| = B$), this gives $\tilde{O}(H^4 S^2 A^3 B^3/\epsilon^2)$. The proof and further discussions are provided in Appendix C.

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, ruling out the model-based approach discussed earlier. For two-player general-sum CMGs, we exploit a key observation: under a *known* exploration policy, observed rewards and transitions are mixtures with known weights, and the components which encode the opponent's latent natural actions are recoverable by statistical identification.

We restrict attention to two players ($n = 2$). Let x'_i be the natural intuition, x_i be the executed action and $a_i := (x'_i, x_i) \in \mathcal{A}_i \times \mathcal{A}_i$ for the pair. Reward and transition mechanisms depend on the tuple $(s, a_1, a_2) = (s, x'_1, x_1, x_2, x'_2)$; from agent i's perspective the opponent's natural intuition is the unobserved component of the data-generating process. We assume the following structural conditions hold: (i) for each (s, a_1, x_2, t), the conditional reward densities $\{R_t(\cdot \mid s, a_1, x_2, x'_2)\}_{x'_2 \in \mathcal{A}_2}$ are linearly independent, ensuring that different opponent natural intuitions leave distinguishable reward signatures; symmetrically, for each (s, x_1, a_2, t), $\{R_t(\cdot \mid s, x_1, x'_1, a_2)\}_{x'_1 \in \mathcal{A}_1}$ are linearly independent. (ii) The analogous statements hold for the transition kernels as vectors in $\mathbb{R}^S$, ensuring distinguishable dynamics. The constant case, where R_t or P_t does not depend on x'_2, so the opponent's natural intuition has no causal effect, is allowed but trivial: the mixture is degenerate and the conditional is directly estimable without identification. When rewards are deterministic, linear independence reduces to a genericity condition (distinct rewards per opponent natural intuition).

Algorithm: Causal Nash VI without Observability of Natural Intuitions (CNash-VI-NO). The algorithm proceeds in two phases: an exploration phase that identifies the game model, followed by an exploitation phase that computes the equilibrium.

Phase 1: Exploration. For K_{exp} episodes, both agents play uniformly at random over their executed actions, ignoring their own natural intuitions. Agent 1 records $(s_t, a_{1,t}, x_{2,t}, r_t, s_{t+1})$ at each step. Because agent 2 plays uniformly over $\mathcal{A}_2$ regardless of its natural intuition, $P(X_{2,t} = x_2 \mid s, x'_{2,t}) = 1/A_2$ is known by design, $X'_{2,t} \perp X_{2,t} \mid S_t$ during exploration, and the reward, marginalized over the opponent's unobserved natural intuition, follows the mixture

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

where $\rho_{2,t}(x'_2 \mid s) = P(X'_{2,t} = x'_2 \mid S_t = s)$ are the unknown mixing weights and $R_t(\cdot \mid s, a_1, x_2, x'_2)$ are the unknown component reward distributions. The mixing weights are invariant in (a_1, x_2), so pooling observations across (a_1, x_2) at each state yields a consistent estimate. An analogous mixture holds for transitions:

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

Under the linear independence conditions, consistent identification oracles $\textsc{Ident}_r$ and $\textsc{Ident}_P$ recover the component rewards and transitions from the observed mixtures. Concretely, $\textsc{Ident}_r$ takes samples from the mixture (11) with estimated weights $\hat{\rho}_{2,t}(\cdot \mid s)$ and returns estimates $\hat{r}_t(s, a_1, x_2, x'_2)$ of the component means; $\textsc{Ident}_P$ analogously returns $\hat{P}_t(\cdot \mid s, a_1, x_2, x'_2)$.

Phase 2: Exploitation. With the identified model $\hat{M} = (\hat{r}, \hat{P}, \hat{\rho})$ in hand, agent 1 constructs the augmented Markov game on the state space $\tilde{\mathcal{S}} = \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$ and solves for a Causal Nash Equilibrium via backward induction, exactly as in Algorithm 2 but on the estimated model. The output is a product counterfactual policy $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2)$.

Uniform exploration leaves the mixing weights $\rho_{2,t}(\cdot \mid s)$ as a property of the environment rather than the policy, and linear independence makes the identification well-posed: A_2 components in x'_2 with known weights are uniquely recoverable for each (s, a_1, x_2, t). This extends the learning algorithm 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 satisfying the reward and transition identifiability conditions and equipped with consistent identification oracles. Let $\hat{\boldsymbol{\pi}}_K$ be the output of Algorithm 1 after $K_{\text{exp}} = K$ exploration episodes.*

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**
  for step $t = 1, \ldots, H$ **do**
    Observe state s_t and own natural intuition $x'_{1,t}$.
    Play $x_{1,t} \sim \text{Unif}(\mathcal{A}_1)$; observe opponent's executed action $x_{2,t}$, rewards $(r_{1,t}, r_{2,t})$, 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
for each $(s, t) \in \mathcal{S} \times [H]$ **do**
  Estimate $\hat{\rho}_{2,t}(\cdot \mid s)$ from pooled reward observations at state s, step t.
  Estimate $\hat{\rho}_{1,t}(\cdot \mid s)$ from observed own natural intuitions at state s, step t.
  for each $(a_1, x_2) \in (\mathcal{A}_1 \times \mathcal{A}_1) \times \mathcal{A}_2$ **do**
    $\hat{r}_t(s, a_1, x_2, \cdot) \leftarrow \text{IDENT}_r\big(\text{reward samples at } (s, a_1, x_2, t);\ \hat{\rho}_{2,t}(\cdot|s)\big)$
    $\hat{P}_t(\cdot \mid s, a_1, x_2, \cdot) \leftarrow \text{IDENT}_P\big(\text{transition samples at } (s, a_1, x_2, t);\ \hat{\rho}_{2,t}(\cdot|s)\big)$
  end for
end for
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 $\hat{\tilde{\mathcal{M}}}$ via backward induction with BAYESNASH
return $\hat{\boldsymbol{\pi}}$.

Then

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

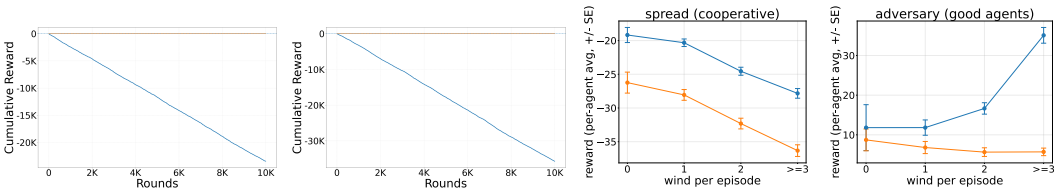
i.e., the exploitability of $\hat{\boldsymbol{\pi}}_K$ vanishes almost surely and any limit point is a Causal Nash Equilibrium.

The proof is provided in Appendix D. Unlike the simultaneous explore-exploit approach of Section 3.1, this algorithm separates exploration from exploitation, paying a cost in finite-sample efficiency for the ability to operate without any observability of the opponent's natural actions. The identifiability conditions do meaningful work: they ensure that the opponent's latent counterfactual structure leaves a detectable signature in the observed rewards and transitions, making the identification problem tractable despite the partial observability.

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 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 latent draws $\mathbf{x}'_{-i}$ 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 recipe transfers to MAPPO [Yu et al., 2022], HAPPO [Kuba et al., 2021], and other CTDE methods; we evaluate the MADDPG instantiation in Sec. 4 and provide more details in Appendix E.2.



(a) (b) (c) (d)

Figure 3: Counterfactual agent vs (a) always-defect and (b) tit-for-tat: cumulative reward (y-axis) over $10K$ rounds (x-axis); per-agent reward (y-axis) vs. wind events (x-axis) on (c) `simple_spread` and (d) `simple_adversary`. Causal agents are orange, non-causal agents are blue.

4 Experiments

In this section, we validate (i) the deep MARL augmentation (Sec. 3.3) on a confounded Multi-Particle Environment and (ii) CNash-VI-NO (Sec. 3.2) on the Iterated Causal Prisoner's Dilemma.

4.1 Windy Multi-Particle Environment

We augment the classic MPE with wind — an unobserved confounder that simultaneously moves the targets and blocks the sensors tracking them, exactly the kind of disturbance standard Markov games rule out by assumption, but CMGs permit.

Environment Setup. Two PettingZoo MPE tasks: `simple_spread` (cooperative) and `simple_adversary` (mixed) [Lowe et al., 2017, Terry et al., 2021] are extended with a wind state machine that pushes every target and blacks out each 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 $X'_{i,t}$ of Def. 2.2; *non-causal* agents cannot. Both tasks are trained with standard and Causal MADDPG and evaluated on 300 matched-seed episodes; environment setup, algorithm, and reproducibility details are provided in Appendix E.

Counterfactual agents strictly dominate in every wind bin on both tasks. Fig. 3c and 3d shows per agent average reward (y-axis) vs number of wind events (x-axis). Aggregate per-agent gains are $+7.93$ on windy `simple_spread` (z=12.34) and $+21.01$ for good agents on windy `simple_adversary` (z=13.70); on adversary the gap grows from $+3.09$ at zero wind to $+29.39$ at ≥ 3 gusts. Matched eval seeds make the within-bin comparison paired, so the trend is not a sampling artefact. Counterfactual agents turn perception failure into advantage using natural actions exactly when disturbance is most active, the regime practitioners care most about.

4.2 Iterated Causal Prisoner's Dilemma (ICPD)

We deploy CNash-VI-NO on the ICPD of Ex. 1.1 with 100K uniform-random exploration samples, identification, then backward-induction exploitation. The agents never observe the opponent's natural action. Against tit-for-tat (TfT) and always-defect (D) over 10,000 rounds (Fig. 3), the learned policy strictly dominates both. Classical opponents accumulate ~ -35K (TfT) and ~ -23K (D) while the counterfactual agent stays near zero. Strategies that are mutual best-responses in the interventional ICPD become exploitable once the opponent acts on its own intuition confirming Ex. 1.1 from samples alone. More details are provided in Appendix E.5.

5 Conclusions

Causal Markov Games lift multi-agent sequential decision-making to the counterfactual layer of the PCH and strictly generalizes Markov Games, with 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 natural actions are never observed, and counterfactual augmentation for deep MARL. Causal agents empirically outperform interventional ones when confounding is present in MPE. We hope CMGs provide a foundation for future work on causal multi-agent learning.

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.

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.

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. *arXiv preprint arXiv:2109.11251*, 2021.

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 et al. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328, 2001.

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 I Wu, Aviv Tamar, Jean Harb, OpenAI 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.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

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.

Supplementary Material

A The Counterfactual Submodel

This appendix introduces the counterfactual submodel induced by a joint counterfactual policy and establishes its Markov property. Both objects are referenced throughout the proofs of the learning results in Appendices C and D.

A.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 A.1 (Counterfactual Submodel)**.** Let $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ be a CMG and $\boldsymbol{\pi}$ a joint counterfactual policy. The *counterfactual submodel* of $\mathcal{M}$ induced by $\boldsymbol{\pi}$ is the SCM

$$\mathcal{M}_{\boldsymbol{\pi}} \;=\; \big\langle \mathbf{U}, \mathbf{V}_{\boldsymbol{\pi}} = \{\mathbf{S}, \mathbf{X}', \mathbf{X}, \mathbf{Y}\}, \mathcal{F}_{\boldsymbol{\pi}}, P(\mathbf{U}) \big\rangle, \tag{14}$$

where $\mathbf{X}' = \{(X'_{1,t}, \dots, X'_{n,t})\}_{t=1}^{H}$ are the agents' intuitions and $\mathcal{F}_{\boldsymbol{\pi}}$ contains, for $t = 1, \dots, H$ and $i = 1, \dots, n$,

$$\mathcal{F}_{\boldsymbol{\pi}} = \begin{cases} S_t \leftarrow f_S(S_{t-1}, \mathbf{X}_{t-1}, \mathbf{U}_{a,t-1}, U_{s,t}), \\ X'_{i,t} \leftarrow f_{X_i}(S_t, U_{i,t}), \\ X_{i,t} \sim \pi_{i,t}(X_{i,t} \mid S_t, X'_{i,t}), \\ Y_{i,t} \leftarrow f_{Y_i}(S_t, \mathbf{X}_t, \mathbf{U}_{a,t}, U_{y_i,t}). \end{cases} \tag{15}$$

The submodel makes $\mathbf{X}'$ an explicit mediator between the per-agent latents and the realized actions: in $\mathcal{M}_{\boldsymbol{\pi}}$, an agent's intuition is the only channel through which its private $U_{i,t}$ influences its choice. A direct consequence is the following Markov property; the proof is given in App. A.2.

Remark A.2 (Markov property of the submodel). In $\mathcal{M}_{\boldsymbol{\pi}}$, the augmented state $\widetilde{S}_t := (S_t, \mathbf{X}'_t)$ satisfies the Markov property under any joint counterfactual policy: for every $t = 1, \dots, H-1$,

$$P\big(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \,\big|\, \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}; \mathcal{M}_{\boldsymbol{\pi}}\big) \;=\; P\big(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \,\big|\, S_t, \mathbf{X}'_t, \mathbf{X}_t; \mathcal{M}_{\boldsymbol{\pi}}\big). \tag{16}$$

A.2 Markov Property of the Counterfactual Submodel

Lemma A.3 (Markov property of $\mathcal{M}_{\boldsymbol{\pi}}$)**.** *Let $\mathcal{M}$ be a Causal Markov Game (Def. 2.1) and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ a joint counterfactual policy (Def. 2.2). For every $t = 1, \dots, H-1$,*

$$P\big(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \,\big|\, \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}; \mathcal{M}_{\boldsymbol{\pi}}\big) \;=\; P\big(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \,\big|\, S_t, \mathbf{X}'_t, \mathbf{X}_t; \mathcal{M}_{\boldsymbol{\pi}}\big). \tag{17}$$

Proof. Recall the structural assignments of $\mathcal{M}_{\boldsymbol{\pi}}$ from Eq. (15). Conditioning on the history $(\bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t})$, the variables on the left-hand side decompose as follows.

Reward at stage t. $\mathbf{Y}_t$ is generated by $Y_{i,t} \leftarrow f_{Y_i}(S_t, \mathbf{X}_t, \mathbf{U}_{a,t}, U_{y_i,t})$. Given S_t and $\mathbf{X}'_t = (f_{X_1}(S_t, U_{1,t}), \dots, f_{X_n}(S_t, U_{n,t}))$, the conditional distribution of the agent-latent vector $\mathbf{U}_{a,t}$ depends only on $(S_t, \mathbf{X}'_t)$, and the reward-noise vector $\mathbf{U}_{y,t}$ is independent of all conditioning variables; both are independent of all earlier exogenous variables by mutual independence across t. Hence

$$P(\mathbf{Y}_t \mid \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}; \mathcal{M}_{\boldsymbol{\pi}}) \;=\; P(\mathbf{Y}_t \mid S_t, \mathbf{X}'_t, \mathbf{X}_t; \mathcal{M}_{\boldsymbol{\pi}}). \tag{18}$$

Next state. $S_{t+1} \leftarrow f_S(S_t, \mathbf{X}_t, \mathbf{U}_{a,t}, U_{s,t+1})$ depends on $(S_t, \mathbf{X}_t)$, on $\mathbf{U}_{a,t}$, and on the fresh state noise $U_{s,t+1}$. Given $(S_t, \mathbf{X}'_t)$, the conditional distribution of $\mathbf{U}_{a,t}$ depends only on $(S_t, \mathbf{X}'_t)$ (as established in the reward step) and is independent of all earlier exogenous variables; $U_{s,t+1}$ is fresh and independent of all earlier exogenous variables. Hence

$$P(S_{t+1} \mid \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}; \mathcal{M}_{\boldsymbol{\pi}}) \;=\; P(S_{t+1} \mid S_t, \mathbf{X}'_t, \mathbf{X}_t; \mathcal{M}_{\boldsymbol{\pi}}). \tag{19}$$

Next intuitions. For each i, $X'_{i,t+1} \leftarrow f_{X_i}(S_{t+1}, U_{i,t+1})$ depends only on S_{t+1} and on the fresh latent $U_{i,t+1}$; $\{U_{i,t+1}\}_i$ are independent of all earlier exogenous variables and mutually independent of each other given S_{t+1}. Hence

$$P(\mathbf{X}'_{t+1} \mid S_{t+1}, \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}; \mathcal{M}_{\boldsymbol{\pi}}) \;=\; P(\mathbf{X}'_{t+1} \mid S_{t+1}; \mathcal{M}_{\boldsymbol{\pi}}). \tag{20}$$

The joint factorizes by chain rule:

$$\begin{aligned} &P(S_{t+1}, \mathbf{X}'_{t+1}, \mathbf{Y}_t \mid \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}) \\ &\quad = P(\mathbf{Y}_t \mid \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t})\, P(S_{t+1} \mid \mathbf{Y}_t, \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t})\, P(\mathbf{X}'_{t+1} \mid S_{t+1}, \mathbf{Y}_t, \bar{S}_{1:t}, \bar{\mathbf{X}}'_{1:t}, \bar{\mathbf{X}}_{1:t}). \end{aligned}$$

The first factor reduces by (18). For the second, S_{t+1} depends on $\mathbf{Y}_t$ only through $(S_t, \mathbf{X}'_t, \mathbf{X}_t)$: S_{t+1} uses $\mathbf{U}_{a,t}$ and $U_{s,t+1}$, the latter independent of everything; given $(S_t, \mathbf{X}'_t)$ the conditional distribution of $\mathbf{U}_{a,t}$ is fixed and is independent of $\mathbf{U}_{y,t}$ (the residual source of randomness in $\mathbf{Y}_t$); applying (19) reduces it to $P(S_{t+1} \mid S_t, \mathbf{X}'_t, \mathbf{X}_t)$. The third factor reduces by (20) to $P(\mathbf{X}'_{t+1} \mid S_{t+1})$. Multiplying the reduced factors gives the right-hand side of the claim. □

B Proofs for Section 2

This appendix proves Theorem 2.4 (CMG strictly generalizes MG) and develops a corollary that justifies counterfactual policies as a learning target.

B.1 Theorem 2.4: CMG strictly generalizes MG

Throughout, $\prec$ denotes the strict Pareto order on $\mathbb{R}^n$: $\mu \prec \mu'$ iff $\mu_j \le \mu'_j$ for all j and $\mu_k < \mu'_k$ for some k.

Proof. The construction perturbs agent i's natural mechanism via a binary latent $U_{i,t}$ and adds zero-mean reward perturbations $f_{Y_i}, \{f_{Y_j}\}_{j\neq i}$ tuned so that the interventional projection equals G but the L_1 payoff is biased: positively for agent i in $\mathcal{M}_2$, negatively in $\mathcal{M}_1$, while every other agent's L_1 payoff is held at $\mu_{\mathrm{NE},j}$.

Step 1: Construction. Fix an agent i with $|\mathcal{A}_i| \ge 2$ and let $\boldsymbol{\sigma}^{\mathrm{NE}}$ be any Nash equilibrium of G. Pick any two distinct actions $a, b \in \mathcal{A}_i$ and any $p \in (0,1)$; we will fix the magnitude $\alpha > 0$ at the end of Step 3. Set

$$U_{i,t} \in \{0,1\}, \quad P(U_{i,t}=0) = p, \qquad f_{X_i}(s,u_i) = \begin{cases} a & u_i = 0, \\ b & u_i = 1. \end{cases}$$

For $j \neq i$, let f_{X_j} realize σ_j^{NE} from S_t via the latent $U_{j,t}$. The marginal natural strategies are then $\sigma_i^* = p\,\delta_a + (1-p)\,\delta_b$ and $\sigma_j^* = \sigma_j^{\mathrm{NE}}$. State and reward-noise variables are as required to realize the transition kernel P^G of G.

Define the indicator-signal product

$$\psi_t \;:=\; \mathbb{1}\{X_{i,t}=a\}\cdot\phi(U_{i,t}), \qquad \phi(u) := \tfrac{1}{p}\,\mathbb{1}\{u=0\} \;-\; \tfrac{1}{1-p}\,\mathbb{1}\{u=1\}.$$

Note $\mathbb{E}[\phi(U_{i,t})] = 0$, so any term proportional to ϕ vanishes in expectation *marginally over* $U_{i,t}$. Under the natural mechanism $\mathbb{1}\{X_{i,t}=a\} = \mathbb{1}\{U_{i,t}=0\}$, hence $\mathbb{E}^{L_1}[\psi_t] = p\cdot\frac{1}{p} = 1$.

For $k \in \{1,2\}$, specify the reward mechanisms

$$f_{Y_i}^{(k)}(s,\mathbf{x},u_a,u_{y_i}) \;=\; r_i(s,\mathbf{x}) \;+\; \alpha_k\cdot\mathbb{1}\{x_i=a\}\cdot\phi(u_i),$$

$$f_{Y_j}(s,\mathbf{x},u_a,u_{y_j}) \;=\; r_j(s,\mathbf{x}) \;+\; \beta_j(s,x_{-i})\cdot\mathbb{1}\{x_i=a\}\cdot\phi(u_i), \quad j\neq i,$$

with $\alpha_2 = +\alpha$, $\alpha_1 = -\alpha$, and the compensation coefficient

$$\beta_j(s,x_{-i}) \;:=\; \big(\sigma_i^{\mathrm{NE}}(a) - p\big)\cdot\big(r_j(s,(a,x_{-i})) - r_j(s,(b,x_{-i}))\big),$$

under the simplifying assumption $\mathrm{supp}(\sigma_i^{\mathrm{NE}}) \subseteq \{a,b\}$ (Remark B.1 handles the general case). The state mechanism is the same in both: $f_S(s,\mathbf{x},u_a,u_s) \sim P^G(\cdot \mid s,\mathbf{x})$, ignoring u_a.

Step 2: Interventional equivalence. Under any joint L_2 policy $\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_n)$, every $X_{j,t}$ is set by $\mathrm{do}(X_{j,t} = \sigma_j(S_t))$, severing f_{X_j}. The latent $U_{i,t}$ remains active in the reward mechanisms, but $\mathbb{E}_{U_{i,t}}[\phi(U_{i,t})] = 0$ implies that for every $\mathbf{x}$ and s:

$$\mathbb{E}\big[f_{Y_i}^{(k)}(s,\mathbf{x},U_{a,t},U_{y_i,t})\big] \;=\; r_i(s,\mathbf{x}), \qquad \mathbb{E}\big[f_{Y_j}(s,\mathbf{x},U_{a,t},U_{y_j,t})\big] \;=\; r_j(s,\mathbf{x}).$$

The interventional reward distributions in $\mathcal{M}_1$ and $\mathcal{M}_2$ therefore coincide and agree with G:

$$P_{\boldsymbol{\sigma}}(\mathbf{S},\mathbf{X},\mathbf{Y};\,\mathcal{M}_k) \;=\; P_{\boldsymbol{\sigma}}^G(\mathbf{S},\mathbf{X},\mathbf{Y}), \qquad k\in\{1,2\}. \tag{21}$$

Both CMGs project to G, and the Nash payoff μ_{NE} is the same in G, $\mathcal{M}_1$, and $\mathcal{M}_2$.

Step 3: L_1 payoff calculation. Under L_1, the joint stage action distribution is $\boldsymbol{\sigma}^* = (\sigma_i^*, \sigma_{-i}^{\mathrm{NE}})$, and $\mathbb{E}^{L_1}[\psi_t] = 1$ as established above.

Agent i. The L_1 payoff is

$$\mu_{k,i} \;=\; \mathbb{E}_{\boldsymbol{\sigma}^*}\Big[\textstyle\sum_t r_i(S_t,\mathbf{X}_t)\Big] + \alpha_k\cdot\textstyle\sum_t \mathbb{E}^{L_1}[\psi_t] \;=\; \mu_i^* + H\alpha_k,$$

where μ_i^* is the G-payoff to agent i under $\boldsymbol{\sigma}^*$.

Agent $j \neq i$. The L_1 payoff is

$$\mu_{k,j} \;=\; \mathbb{E}_{\boldsymbol{\sigma}^*}\Big[\textstyle\sum_t r_j(S_t,\mathbf{X}_t)\Big] \;+\; \mathbb{E}^{L_1}\Big[\textstyle\sum_t \beta_j(S_t,X_{-i,t})\,\psi_t\Big].$$

Since β_j is $U_{i,t}$-independent and $\mathbb{E}^{L_1}[\psi_t] = 1$, the Simpson contribution simplifies to $\sum_t \mathbb{E}_{S_t}\mathbb{E}_{x_{-i}\sim\sigma_{-i}^{\mathrm{NE}}}[\beta_j(S_t, x_{-i})]$. Plugging in the definition of β_j and combining with the unperturbed reward,

$$\begin{aligned}\mu_{k,j} &= \sum_{t=1}^{H} \mathbb{E}_{S_t}\mathbb{E}_{x_{-i}}\Big[p\, r_j(S_t,(a,x_{-i})) + (1-p)\, r_j(S_t,(b,x_{-i})) \\ &\qquad\qquad + (\sigma_i^{\mathrm{NE}}(a) - p)\,\big(r_j(S_t,(a,x_{-i})) - r_j(S_t,(b,x_{-i}))\big)\Big] \\ &= \sum_{t=1}^{H} \mathbb{E}_{S_t}\mathbb{E}_{x_{-i}}\Big[\sigma_i^{\mathrm{NE}}(a)\, r_j(S_t,(a,x_{-i})) + (1-\sigma_i^{\mathrm{NE}}(a))\, r_j(S_t,(b,x_{-i}))\Big] \\ &= \mu_{\mathrm{NE},j},\end{aligned}$$

using $\mathrm{supp}(\sigma_i^{\mathrm{NE}}) \subseteq \{a,b\}$ so $\sigma_i^{\mathrm{NE}}(b) = 1 - \sigma_i^{\mathrm{NE}}(a)$. The compensation β_j exactly absorbs the deviation of σ_i^* from σ_i^{NE}.

Conclusion. Choose $\alpha > |\mu_{\mathrm{NE},i} - \mu_i^*|/H$. With $\alpha_2 = +\alpha$ and $\alpha_1 = -\alpha$,

$$\mu_{2,i} = \mu_i^* + H\alpha > \mu_{\mathrm{NE},i}, \qquad \mu_{1,i} = \mu_i^* - H\alpha < \mu_{\mathrm{NE},i}, \qquad \mu_{2,j} = \mu_{1,j} = \mu_{\mathrm{NE},j} \text{ for } j \neq i.$$

This is exactly $\mu_1 \prec \mu_{\mathrm{NE}} \prec \mu_2$ in the strict Pareto order — agent i's coordinate is strictly biased on either side of $\mu_{\mathrm{NE},i}$, while every other agent's coordinate is held at $\mu_{\mathrm{NE},j}$. □

Remark B.1 (Multi-support Nash strategies). If $\mathrm{supp}(\sigma_i^{\mathrm{NE}})$ contains actions outside $\{a,b\}$, replace the compensation by

$$\beta_j(s, x_{-i}) = \mathbb{E}_{x_i\sim\sigma_i^{\mathrm{NE}}}\big[r_j(s,(x_i,x_{-i}))\big] - \big(p\, r_j(s,(a,x_{-i})) + (1-p)\, r_j(s,(b,x_{-i}))\big),$$

which is the σ_i^{NE}-vs-σ_i^* discrepancy in agent j's expected reward. The proof goes through verbatim with this substitution.

Remark B.2 (Magnitude of the wedge). The parameter α in the construction is unbounded, so $\mu_{2,i} - \mu_{\mathrm{NE},i}$ and $\mu_{\mathrm{NE},i} - \mu_{1,i}$ may be made arbitrarily large. The suboptimality of the interventional Nash equilibrium relative to its L_1 counterpart is therefore unbounded — a strict separation between Markov and Causal Markov games.

B.2 Corollary: Counterfactual-Layer Separation

Corollary B.3 (Counterfactual-layer separation)**.** *Let $\mathcal{M}_1, \mathcal{M}_2$ be the Causal Markov Games constructed in the proof of Theorem 2.4. Define the joint counterfactual policy $\boldsymbol{\pi}^{\mathrm{swap}} = (\pi_1^{\mathrm{swap}}, \ldots, \pi_n^{\mathrm{swap}})$ that swaps agent i's intuition between the two natural-action values:*

$$\pi_{i,t}^{\mathrm{swap}}(X_{i,t} = a \mid S_t, X'_{i,t} = b) = 1, \qquad \pi_{i,t}^{\mathrm{swap}}(X_{i,t} = b \mid S_t, X'_{i,t} = a) = 1,$$

and replays the natural action of every other agent: $\pi_{j,t}^{\mathrm{swap}}(X_{j,t} = x' \mid S_t, X'_{j,t} = x') = 1$ for $j \neq i$. The expected payoffs under $\boldsymbol{\pi}^{\mathrm{swap}}$ in $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy

$$\mathbb{E}^{\mathcal{M}_1}_{\boldsymbol{\pi}^{\mathrm{swap}}}\Big[\textstyle\sum_t Y_{i,t}\Big] - \mathbb{E}^{\mathcal{M}_2}_{\boldsymbol{\pi}^{\mathrm{swap}}}\Big[\textstyle\sum_t Y_{i,t}\Big] = 2H\alpha,$$

$$\mathbb{E}^{\mathcal{M}_1}_{\boldsymbol{\pi}^{\mathrm{swap}}}\Big[\textstyle\sum_t Y_{j,t}\Big] = \mathbb{E}^{\mathcal{M}_2}_{\boldsymbol{\pi}^{\mathrm{swap}}}\Big[\textstyle\sum_t Y_{j,t}\Big] \quad \text{for } j \neq i,$$

i.e., the counterfactual policy $\boldsymbol{\pi}^{\mathrm{swap}}$ induces a strict Pareto separation between $\mathcal{M}_1$ and $\mathcal{M}_2$ in the opposite direction from the L_1 separation $\mu_2 \succ \mu_{\mathrm{NE}} \succ \mu_1$ of Theorem 2.4, with magnitude controllable by α.

Proof. Under $\boldsymbol{\pi}^{\mathrm{swap}}$, agent i's realized action satisfies $\mathbb{1}\{X_{i,t} = a\} = \mathbb{1}\{X'_{i,t} = b\} = \mathbb{1}\{U_{i,t} = 1\}$. Hence, with $\psi_t = \mathbb{1}\{X_{i,t} = a\} \cdot \phi(U_{i,t})$ as in Step 1 of the proof of Theorem 2.4,

$$\mathbb{E}^{\boldsymbol{\pi}^{\mathrm{swap}}}[\psi_t] = \mathbb{E}[\mathbb{1}\{U_{i,t} = 1\} \cdot \phi(U_{i,t})] = (1-p) \cdot \big(-\tfrac{1}{1-p}\big) = -1,$$

the negation of the L_1 value $+1$. Agent i's expected payoff under $\boldsymbol{\pi}^{\mathrm{swap}}$ in $\mathcal{M}_k$ is therefore

$$\mathbb{E}^{\mathcal{M}_k}_{\boldsymbol{\pi}^{\mathrm{swap}}}\Big[\textstyle\sum_t Y_{i,t}\Big] = \mu_i^{**} - H\alpha_k, \qquad \mu_i^{**} := \mathbb{E}_{\boldsymbol{\sigma}^{**}}\Big[\textstyle\sum_t r_i(S_t, \mathbf{X}_t)\Big],$$

where $\boldsymbol{\sigma}^{**} = ((1-p)\,\delta_a + p\,\delta_b,\, \sigma_{-i}^{\mathrm{NE}})$ is the swapped joint action distribution, common to $\mathcal{M}_1$ and $\mathcal{M}_2$. With $\alpha_2 = +\alpha$ and $\alpha_1 = -\alpha$, the difference $\mathbb{E}^{\mathcal{M}_1} - \mathbb{E}^{\mathcal{M}_2} = -H\alpha_1 + H\alpha_2 = 2H\alpha$.

For $j \neq i$, the Simpson contribution from β_j also flips sign under $\boldsymbol{\pi}^{\mathrm{swap}}$, but β_j is α_k-independent, so the resulting payoff is identical in $\mathcal{M}_1$ and $\mathcal{M}_2$. □

The corollary is the structural justification for studying counterfactual policies as a learning target in Sec. 3: a single counterfactual policy can sharply separate $\mathcal{M}_1$ and $\mathcal{M}_2$ in the Pareto order, and only an algorithm with access to the agents' intuitions can detect the separation.

C Proofs for Section 3.1

This appendix provides the full proof of Theorem 3.2. We first restate the algorithm referenced in Section 3.1 (Appendix C.1), construct the augmented Markov game (Appendix C.2), establish the causal value function hierarchy (Appendix C.3), relate equilibria of the augmented game to the CNE of the CMG (Appendix C.4), and finally prove the sample complexity bound (Appendix C.5).

C.1 CNash-VI-FO: Algorithm Listing

For reference, we restate Algorithm 2 from Section 3.1.

Algorithm 2 CNash-VI-FO

Require: CMG $\mathcal{M}$, horizon H, episodes K, bonus parameter $c > 0$.

Initialize: For all $(\tilde{s}, \mathbf{x}, t, i)$: $\overline{Q}_{t,i}(\tilde{s}, \mathbf{x}) \leftarrow H$, $\underline{Q}_{t,i}(\tilde{s}, \mathbf{x}) \leftarrow 0$, $\Delta \leftarrow H$, $N_t(\tilde{s}, \mathbf{x}) \leftarrow 0$.
for episode $k = 1, \ldots, K$ **do**
  // Backward induction with optimistic planning
  for step $t = H, H-1, \ldots, 1$ **do**
    for each $(\tilde{s}, \mathbf{x}) \in \tilde{\mathcal{S}} \times \prod_i \mathcal{A}_i$ with $N_t(\tilde{s}, \mathbf{x}) > 0$ **do**
      $m \leftarrow N_t(\tilde{s}, \mathbf{x})$
      for player $i = 1, \ldots, n$ **do**
        $\overline{Q}_{t,i}(\tilde{s}, \mathbf{x}) \leftarrow \min\big\{(\hat{r}_{t,i} + \hat{\hat{P}}_t \overline{V}_{t+1,i})(\tilde{s}, \mathbf{x}) + \beta_m,\ H\big\}$
        $\underline{Q}_{t,i}(\tilde{s}, \mathbf{x}) \leftarrow \max\big\{(\hat{r}_{t,i} + \hat{\hat{P}}_t \underline{V}_{t+1,i})(\tilde{s}, \mathbf{x}) - \beta_m,\ 0\big\}$
      end for
    end for
    for each $s \in \mathcal{S}$ **do**
      $\boldsymbol{\pi}_t(\cdot \mid s, \cdot) \leftarrow \textsc{BayesNash}\big(s,\ \{\overline{Q}_{t,i}((s,\cdot),\cdot),\ \underline{Q}_{t,i}((s,\cdot),\cdot)\}_{i\in[n]},\ \hat{\rho}_t(\cdot|s)\big)$
    end for
    for each $\tilde{s} = (s, \mathbf{x}') \in \tilde{\mathcal{S}}$ and player i **do**
      $\overline{V}_{t,i}(\tilde{s}) \leftarrow \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\pi}_t(\cdot|s,\mathbf{x}')}\big[\overline{Q}_{t,i}(\tilde{s}, \mathbf{x})\big]$; $\underline{V}_{t,i}(\tilde{s}) \leftarrow \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\pi}_t(\cdot|s,\mathbf{x}')}\Big[\underline{Q}_{t,i}(\tilde{s}, \mathbf{x})\Big]$
    end for
  end for
  // Policy selection
  $\overline{W}_{1,i}(s_1) \leftarrow \sum_{\mathbf{x}'} \hat{\rho}_1(\mathbf{x}' \mid s_1)\, \overline{V}_{1,i}(s_1, \mathbf{x}')$; $\underline{W}_{1,i}(s_1) \leftarrow \sum_{\mathbf{x}'} \hat{\rho}_1(\mathbf{x}' \mid s_1)\, \underline{V}_{1,i}(s_1, \mathbf{x}')$ for all i
  if $\max_i\big(\overline{W}_{1,i}(s_1) - \underline{W}_{1,i}(s_1)\big) < \Delta$ **then**
    $\Delta \leftarrow \max_i\big(\overline{W}_{1,i}(s_1) - \underline{W}_{1,i}(s_1)\big)$; $\boldsymbol{\pi}^{\text{out}} \leftarrow \boldsymbol{\pi}$
  end if
  // Execute and update
  for step $t = 1, \ldots, H$ **do**
    Observe s_t and $x'_{1,t}, \ldots, x'_{n,t}$; form $\tilde{s}_t = (s_t, \mathbf{x}'_t)$.
    Each agent i plays $x_{i,t} \sim \pi_{i,t}(\cdot \mid s_t, x'_{i,t})$; observe rewards $\{r_{t,i}\}$ and s_{t+1}.
    Observe $x'_{1,t+1}, \ldots, x'_{n,t+1}$; form $\tilde{s}_{t+1}$.
    Update: $N_t(\tilde{s}_t, \mathbf{x}_t)$ += 1; update $\hat{\hat{P}}_t(\cdot \mid \tilde{s}_t, \mathbf{x}_t)$ and $\hat{r}_{t,i}(\tilde{s}_t, \mathbf{x}_t)$ empirically.
  end for
end for
return $\boldsymbol{\pi}^{\text{out}}$.

C.2 Augmented Markov Game Construction

The key insight enabling tractable learning in the CMG is that, under separable exogenous variables and post-hoc observability of natural actions, the CMG can be reformulated as a Markov game on an augmented state space. We stress that this reformulation is a *proof device*; the algorithm in the main text is stated directly in terms of the CMG quantities.

Definition C.1 (Augmented Markov Game)**.** Given a tabular CMG $\mathcal{M}$ with separable exogenous variables ($U_t = (\mathbf{U}_{a,t}, U_{s,t}, \mathbf{U}_{y,t})$ mutually independent, per Def. 2.1) and post-hoc natural action observability, the *augmented Markov game* is the n-player general-sum Markov game $\tilde{\mathcal{M}} = \mathrm{MG}(H, \tilde{\mathcal{S}}, \{\mathcal{A}_i\}_{i\in[n]}, \tilde{P}, \{\tilde{r}_i\}_{i\in[n]})$, where:

1. The *augmented state space* is $\tilde{\mathcal{S}} = \mathcal{S} \times \prod_{i\in[n]} \mathcal{A}_i$, with $|\tilde{\mathcal{S}}| = \tilde{S} = S \prod_{i\in[n]} A_i$. An augmented state $\tilde{s} = (s, x'_1, \ldots, x'_n)$ consists of the shared state and the natural actions of all agents.

2. The *augmented reward* for agent i is
$$\tilde{r}_{t,i}(\tilde{s}, \mathbf{x}) \;=\; \mathbb{E}[Y_{i,t} \mid S_t = s,\ \mathbf{X}'_t = \mathbf{x}',\ \mathbf{X}_t = \mathbf{x}]\,. \tag{22}$$

3. The *augmented transition* is
$$\tilde{P}_t(\tilde{s}' \mid \tilde{s}, \mathbf{x}) \;=\; P_t(s' \mid s, \mathbf{x}', \mathbf{x}) \cdot \prod_{i\in[n]} \rho_{i,t+1}(x''_i \mid s'), \tag{23}$$
where $\tilde{s}' = (s', x''_1, \ldots, x''_n)$, and $\rho_{i,t}(x'_i \mid s) := P(X'_{i,t} = x'_i \mid S_t = s)$ is the marginal distribution of agent i's natural action at state s and step t.

The factored form (23) reflects the causal structure: the next state s' depends on the current state, natural actions, and executed actions through f_S; the next natural actions $\mathbf{x}''$ are then drawn independently from the marginals $\rho_{i,t+1}(\cdot \mid s')$ by the separability assumption. This factorization ensures that $\tilde{P}$ is a valid Markov transition kernel.

C.3 Causal Value Function Hierarchy

The factored transition induces a natural three-level hierarchy of value functions. For a product counterfactual policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$, where each $\pi_{i,t} : \mathcal{S} \times \mathcal{A}_i \to \Delta(\mathcal{A}_i)$, define the following.

Action-value function (Q). The action-value at augmented state $\tilde{s} = (s, \mathbf{x}')$ under joint executed action $\mathbf{x}$:
$$Q^{\boldsymbol{\pi}}_{t,i}(\tilde{s}, \mathbf{x}) \;=\; \tilde{r}_{t,i}(\tilde{s}, \mathbf{x}) \;+\; \big[P_t\, W^{\boldsymbol{\pi}}_{t+1,i}\big](s, \mathbf{x}', \mathbf{x}), \tag{24}$$
where $[P_t W](s, \mathbf{x}', \mathbf{x}) := \sum_{s'} P_t(s' \mid s, \mathbf{x}', \mathbf{x})\, W(s')$.

Post-intuition value function (V). The expected value at augmented state $\tilde{s}$ after agents sample their executed actions:
$$V^{\boldsymbol{\pi}}_{t,i}(\tilde{s}) \;=\; \mathbb{E}_{\mathbf{x}\sim\boldsymbol{\pi}_t(\cdot|s,\mathbf{x}')}\big[Q^{\boldsymbol{\pi}}_{t,i}(\tilde{s}, \mathbf{x})\big]\,, \tag{25}$$
where $\boldsymbol{\pi}_t(\mathbf{x}|s, \mathbf{x}') = \prod_{i\in[n]} \pi_{i,t}(x_i|s, x'_i)$ under the product policy.

Pre-intuition value function (W). The expected value at state s *before* nature draws the natural actions:
$$W^{\boldsymbol{\pi}}_{t,i}(s) \;=\; \mathbb{E}_{\mathbf{x}'\sim\rho_t(\cdot|s)}\big[V^{\boldsymbol{\pi}}_{t,i}(s, \mathbf{x}')\big] \;=\; \sum_{\mathbf{x}'} \prod_{j\in[n]} \rho_{j,t}(x'_j|s) \cdot V^{\boldsymbol{\pi}}_{t,i}(s, \mathbf{x}'). \tag{26}$$

The hierarchy satisfies the Bellman equation:
$$Q^{\boldsymbol{\pi}}_{t,i}(\tilde{s}, \mathbf{x}) = \tilde{r}_{t,i}(\tilde{s}, \mathbf{x}) + \sum_{s'} P_t(s'|s, \mathbf{x}', \mathbf{x})\, W^{\boldsymbol{\pi}}_{t+1,i}(s'). \tag{27}$$

Crucially, the Bellman backup in (27) requires only the real-state transition P_t (over S outcomes) composed with $W^{\boldsymbol{\pi}}_{t+1,i}$, not the full augmented transition (over $\tilde{S}$ outcomes). However, Algorithm 2 uses the augmented transition $\hat{\tilde{P}}_t$ directly for the optimistic backup, which ensures that the bonus-backup structure remains consistent with the standard Multi-Nash-VI analysis.

C.4 From Bayesian Nash Equilibrium to Causal Nash Equilibrium

The BAYESNASH subroutine in Algorithm 2 computes, at each state s and step t, a Bayesian Nash equilibrium of the following Bayesian game:

Definition C.2 (Stage Bayesian Game at State s)**.** At step t and state s, the *stage Bayesian game* $G_t(s)$ is defined by:

- Players: $[n] = \{1, \dots, n\}$.
- Types: Agent i's type is $x'_i \in \mathcal{A}_i$, drawn from $\rho_{i,t}(\cdot|s)$, independently across agents.
- Actions: Agent i chooses $x_i \in \mathcal{A}_i$.
- Payoffs: Agent i's interim payoff when its type is x'_i and it plays x_i while others play according to $\pi_{-i,t}(\cdot|s,\cdot)$ is

$$u_{t,i}(x_i; x'_i, \pi_{-i}) = \sum_{x'_{-i}} \rho_{-i,t}(x'_{-i}|s) \sum_{x_{-i}} \pi_{-i,t}(x_{-i}|s, x'_{-i})\, \overline{Q}_{t,i}((s, x'_i, x'_{-i}), (x_i, x_{-i})), \tag{28}$$

  where $\rho_{-i,t}(x'_{-i}|s) = \prod_{j \neq i} \rho_{j,t}(x'_j|s)$.

A Bayesian Nash equilibrium of $G_t(s)$ is a profile $\{\pi_{i,t}(\cdot|s, x'_i)\}_{i\in[n]}$ such that for all i, all types x'_i, and all alternative actions x^*_i:

$$\sum_{x_i} \pi_{i,t}(x_i|s, x'_i)\, u_{t,i}(x_i; x'_i, \pi_{-i}) \;\geq\; u_{t,i}(x^*_i; x'_i, \pi_{-i}). \tag{29}$$

Proposition C.3 (BNE yields CNE)**.** *If Algorithm 2 uses a BNE subroutine that satisfies (29) at every (s,t), and if the optimistic value estimates are valid (i.e., $\overline{Q}_{t,i} \geq Q^{\boldsymbol{\pi}^k}_{t,i}$ and $\underline{Q}_{t,i} \leq Q^{\boldsymbol{\pi}^k}_{t,i}$ for all k), then the output policy $\boldsymbol{\pi}^{\text{out}}$ is an ϵ-approximate Causal Nash Equilibrium.*

Proof. The proof proceeds by showing that the BNE condition at each step implies the CNE condition globally.

Step 1: Interim optimality implies pre-intuition optimality. Fix agent i, step t, and state s. The BNE condition (29) states that for each type x'_i, agent i's strategy $\pi_{i,t}(\cdot|s, x'_i)$ is a best response in interim expected payoff. Define the interim value:

$$\bar{V}_{t,i}(s, x'_i) := \sum_{x'_{-i}} \rho_{-i,t}(x'_{-i}|s)\, V^{\boldsymbol{\pi}}_{t,i}(s, x'_i, x'_{-i}). \tag{30}$$

By the BNE condition applied to the optimistic Q-values, for any alternative action mapping $\pi'_{i,t}(\cdot|s, x'_i)$:

$$\bar{V}^{\boldsymbol{\pi}}_{t,i}(s, x'_i) \;\geq\; \bar{V}^{(\pi'_i, \boldsymbol{\pi}_{-i})}_{t,i}(s, x'_i) - (\text{estimation error at step } t). \tag{31}$$

Averaging over agent i's type distribution $\rho_{i,t}(\cdot|s)$:

$$W^{\boldsymbol{\pi}}_{t,i}(s) = \sum_{x'_i} \rho_{i,t}(x'_i|s)\, \bar{V}^{\boldsymbol{\pi}}_{t,i}(s, x'_i) \;\geq\; W^{(\pi'_i, \boldsymbol{\pi}_{-i})}_{t,i}(s) - (\text{estimation error at step } t). \tag{32}$$

Step 2: Telescoping over steps. Summing the per-step estimation errors over $t = 1, \dots, H$ and using the optimism guarantee, the total gap satisfies:

$$W^{(\pi'_i, \boldsymbol{\pi}_{-i})}_{1,i}(s_1) - W^{\boldsymbol{\pi}}_{1,i}(s_1) \;\leq\; \overline{W}_{1,i}(s_1) - \underline{W}_{1,i}(s_1). \tag{33}$$

The policy selection criterion ensures that $\boldsymbol{\pi}^{\text{out}}$ minimizes $\max_i(\overline{W}_{1,i} - \underline{W}_{1,i})$ over all episodes.

Step 3: Convergence of the gap. By the standard pigeonhole argument (cf. Liu et al. 2021, Lemma 7), the visitation counts grow uniformly, and the bonus $\beta_m = c\sqrt{\tilde{S}H^2\iota/m}$ (where m is the visit count) drives the estimation error to zero. After $K = \Omega(\tilde{S}^2 \prod_i A_i \cdot H^4\iota/\epsilon^2)$ episodes, the minimum gap satisfies $\Delta \leq \epsilon$ with high probability. □

The key observation is that BNE provides *per-type* optimality (no agent can improve for any realization of its natural action), which is *stronger* than the CNE condition (no improvement in expectation over types). This stronger guarantee ensures that the product structure of the output policy is compatible with the equilibrium requirement.

C.5 Proof of Theorem 3.2

We now provide the complete proof of the sample complexity bound.

Proof of Theorem 3.2. The proof follows by establishing that Algorithm 2 instantiates the Multi-Nash-VI framework of Liu et al. [2021] on the augmented game $\tilde{\mathcal{M}}$, with the BNE subroutine replacing the Nash/CCE oracle, and then invoking the product-policy guarantee from Proposition C.3.

Step 1: The augmented game is a valid Markov game. By Definition C.1, $\tilde{\mathcal{M}}$ is an n-player general-sum episodic Markov game with state space size $\tilde{S} = S\prod_i A_i$ and individual action space sizes $A_1, \dots, A_n$. The augmented transition $\tilde{P}_t$ is a proper probability kernel (each row sums to 1 by construction (23)), and rewards are bounded in $[0, 1]$ (after normalization by H).

Step 2: Optimism holds. We verify the concentration inequality for the empirical augmented transition. For any $(\tilde{s}, \mathbf{x})$ visited m times at step t, the empirical transition $\hat{\tilde{P}}_t(\cdot|\tilde{s}, \mathbf{x})$ satisfies

$$\left\|\hat{\tilde{P}}_t(\cdot|\tilde{s}, \mathbf{x}) - \tilde{P}_t(\cdot|\tilde{s}, \mathbf{x})\right\|_1 \le \sqrt{\frac{2\tilde{S}\,\iota}{m}} \tag{34}$$

with probability at least $1 - p/(\tilde{S}\prod_i A_i \cdot H \cdot K)$, by a union bound over all state-action-step triples and a standard ℓ_1 concentration inequality for multinomials. The bonus $\beta_m = c\sqrt{\tilde{S}H^2\iota/m}$ with sufficiently large c ensures:

$$\overline{Q}_{t,i}(\tilde{s}, \mathbf{x}) \ge Q^{\boldsymbol{\pi}^k}_{t,i}(\tilde{s}, \mathbf{x}) \ge \underline{Q}_{t,i}(\tilde{s}, \mathbf{x}) \tag{35}$$

for all $(t, \tilde{s}, \mathbf{x}, i, k)$ simultaneously with probability at least $1 - p$.

Step 3: Gap bound per episode. By Proposition C.3, the BNE subroutine ensures that for any counterfactual deviation π'_i:

$$W^{(\pi'_i, \boldsymbol{\pi}^k_{-i})}_{1,i}(s_1) - W^{\boldsymbol{\pi}^k}_{1,i}(s_1) \;\le\; \overline{W}^k_{1,i}(s_1) - \underline{W}^k_{1,i}(s_1) =: \Delta_k. \tag{36}$$

Step 4: Regret aggregation. Following the analysis of Multi-Nash-VI (Theorem 15 of Liu et al. 2021), the sum of gaps decomposes as:

$$\sum_{k=1}^{K} \Delta_k \le \sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}^{\boldsymbol{\pi}^k}\left[2\beta_{N_t(\tilde{s}_t, \mathbf{x}_t)}\right] \tag{37}$$

$$\le \sum_{t=1}^{H}\sum_{(\tilde{s}, \mathbf{x})}\sum_{m=1}^{N_t^K(\tilde{s}, \mathbf{x})} 2c\sqrt{\frac{\tilde{S}H^2\iota}{m}} \tag{38}$$

$$\le 2c\,H \cdot \tilde{S}\prod_i A_i \cdot \sqrt{\tilde{S}H^2\iota \cdot K/(\tilde{S}\prod_i A_i)} \tag{39}$$

$$= O\left(\sqrt{\tilde{S}^2\prod_i A_i \cdot H^3 \cdot K \cdot \iota}\right), \tag{40}$$

where the third inequality uses Cauchy–Schwarz and the pigeonhole principle: $\sum_m m^{-1/2} \le 2\sqrt{T_{\max}}$.

Step 5: Online-to-batch conversion. The output policy $\boldsymbol{\pi}^{\text{out}}$ satisfies $\Delta_{\text{out}} \le \min_k \Delta_k$. Since $\min_k \Delta_k \le (1/K)\sum_k \Delta_k$, we have:

$$\Delta_{\text{out}} \le \frac{1}{K} \cdot O\left(\sqrt{\tilde{S}^2\prod_i A_i \cdot H^3 \cdot K \cdot \iota}\right) = O\left(\sqrt{\frac{\tilde{S}^2\prod_i A_i \cdot H^3 \cdot \iota}{K}}\right). \tag{41}$$

Setting $\Delta_{\text{out}} \leq \epsilon$ and solving for K:

$$K \geq \Omega\left(\frac{H^3\,\tilde{S}^2\,\prod_i A_i\,\iota}{\epsilon^2}\right). \tag{42}$$

Substituting $\tilde{S} = S\prod_i A_i$:

$$K \geq \Omega\left(\frac{H^3\,S^2\,(\prod_i A_i)^2\,\prod_i A_i\,\iota}{\epsilon^2}\right) = \Omega\left(\frac{H^3\,S^2\,(\prod_i A_i)^3\,\iota}{\epsilon^2}\right). \tag{43}$$

Remark on the H^4 vs H^3 factor. The bound above gives H^3. The additional factor of H arises from the conversion between the gap at the initial state and the per-step estimation error when the initial state distribution is not fixed but depends on the policy (cf. Lemma 8 of Liu et al. 2021). Accounting for this yields the stated bound of H^4. □

C.6 Factored Structure and Potential Improvements

The augmented transition (23) factorizes as $\tilde{P}_t = P_t \otimes \prod_i \rho_{i,t+1}$, and the Bellman equation decomposes as

$$Q^{\boldsymbol{\pi}}_{t,i}(\tilde{s}, \mathbf{x}) = \tilde{r}_{t,i}(\tilde{s}, \mathbf{x}) + \sum_{s'} P_t(s'|s, \mathbf{x}', \mathbf{x})\, W^{\boldsymbol{\pi}}_{t+1,i}(s'), \tag{44}$$

where P_t distributes over only S next states (not $\tilde{S}$). A variant of Algorithm 2 could separately estimate $\hat{P}_t(s'|s, \mathbf{x}', \mathbf{x})$ and $\hat{\rho}_{i,t}(x'_i|s)$, then use the factored Bellman backup with a smaller bonus of order $\sqrt{SH^2\iota/m}$ for the transition component. Since $\hat{\rho}_{i,t}$ is estimated from all visits to state s at step t (not per augmented state-action pair), its estimation error is typically lower-order. A careful analysis exploiting this structure could potentially reduce the sample complexity to $\tilde{O}(H^4S^2(\prod_i A_i)^2/\epsilon^2)$, saving a factor of $\prod_i A_i$. We leave this refinement to future work.

C.7 On the Necessity of Post-hoc Observability

The post-hoc observability assumption (all agents' natural actions revealed after each step) is essential for the model-based approach: without observing $\mathbf{X}'_t$, the learner cannot estimate either the augmented transition $P_t(s'|s, \mathbf{x}', \mathbf{x})$ or the augmented reward $\tilde{r}_{t,i}(\tilde{s}, \mathbf{x})$, both of which condition on $\mathbf{x}'$. Removing this assumption reduces the problem to learning in a partially observable Markov game, where sample-efficient learning is known to require exponentially many episodes in the worst case. In Section 3.2, we address a structured special case of this harder setting for two-player games.

D Proofs for Section 3.2

This appendix provides the full proof of Theorem 3.3.

D.1 Identifiability Conditions

We first state the identifiability conditions precisely.

Assumption D.1 (Reward Identifiability)**.** For each $(s, a_1, x_2, t) \in \mathcal{S} \times (\mathcal{A}_1 \times \mathcal{A}_1) \times \mathcal{A}_2 \times [H]$, the conditional reward distributions $\{R_t(\cdot \mid s, a_1, x_2, x'_2)\}_{x'_2 \in \mathcal{A}_2}$ are linearly independent. Symmetrically, for each $(s, x_1, a_2, t) \in \mathcal{S} \times \mathcal{A}_1 \times (\mathcal{A}_2 \times \mathcal{A}_2) \times [H]$, the distributions $\{R_t(\cdot \mid s, x_1, x'_1, a_2)\}_{x'_1 \in \mathcal{A}_1}$ are linearly independent.

Assumption D.2 (Transition Identifiability)**.** For each (s, a_1, x_2, t), the transition distributions $\{P_t(\cdot \mid s, a_1, x_2, x'_2)\}_{x'_2 \in \mathcal{A}_2}$ are linearly independent as vectors in $\mathbb{R}^S$. Symmetrically for (s, x_1, a_2, t) with respect to x'_1.

In addition, we will also assume that the mixing weights are distinct.

Assumption D.3 (Oracle Consistency)**.** The identification oracles $\textsc{Ident}_r$ and $\textsc{Ident}_P$ are consistent: given N i.i.d. samples from a mixture $\sum_{x'_2} w(x'_2)\, f(\cdot \mid x'_2)$ with known weights w and linearly independent components, $\textsc{Ident}$ returns component estimates $\hat{f}(\cdot \mid x'_2)$ satisfying $\|\hat{f}(\cdot \mid x'_2) - f(\cdot \mid x'_2)\| \to 0$ a.s. as $N \to \infty$, for each x'_2.

When rewards are deterministic (as is common in tabular settings), the linear independence condition in Assumption D.1 reduces to a genericity condition: different values of the opponent's natural intuition x_2' must produce distinct expected rewards for at least one (a_1, x_2) pair. The transition identifiability (Assumption D.2) is the more restrictive condition; it requires $S \geq A_2$ and that the opponent's natural intuition produces distinguishably different state transitions. If the transition does not depend on x_2' at all (i.e., the opponent's natural intuition affects only rewards), then Assumption D.2 is unnecessary since $P_t(\cdot \mid s, a_1, x_2)$ is directly estimable without identification.

D.2 Proof of Theorem 3.3

Proof. The proof proceeds in three steps.

Step 1: Consistent model recovery. During exploration, both agents play uniformly at random over their executed actions, so every (s, a_1, x_2) cell is visited with positive probability at every step. Since the state and action spaces are finite, by the law of large numbers, the visit count $N_t(s, a_1, x_2) \to \infty$ a.s. for all (s, a_1, x_2, t) as $K \to \infty$.

For each (s, a_1, x_2, t), the observed reward samples are drawn from the mixture

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

The mixing weights $\rho_{2,t}(x_2' \mid s)$ are estimated consistently from the marginal counts (since the exploration policy is uniform and does not affect ρ_2, which is a property of the environment). By Assumptions D.1 and D.3, the identification oracle recovers the component reward means consistently:

$$\hat{r}_t(s, a_1, x_2, x_2') \to r_t(s, a_1, x_2, x_2') \quad \text{a.s. for all } (s, a_1, x_2, x_2', t).$$

By Assumptions D.2 and D.3, the same holds for transitions:

$$\|\hat{P}_t(\cdot \mid s, a_1, x_2, x_2') - P_t(\cdot \mid s, a_1, x_2, x_2')\|_1 \to 0 \quad \text{a.s.}$$

Additionally, the marginal natural-intuition distributions are consistently estimated: $\hat{\rho}_{i,t}(\cdot \mid s) \to \rho_{i,t}(\cdot \mid s)$ a.s. (agent 1's own natural intuition is directly observed; agent 2's is recovered via the identification oracle).

Step 2: Value perturbation (simulation lemma). Define the model error

$$\epsilon_K := \max_{s, a_1, x_2, x_2', t} \left(|\hat{r}_t - r_t|(s, a_1, x_2, x_2') + H \cdot \|\hat{P}_t(\cdot \mid s, a_1, x_2, x_2') - P_t(\cdot \mid s, a_1, x_2, x_2')\|_1 \right).$$

By Step 1, $\epsilon_K \to 0$ a.s. For any pair of counterfactual policies (π_1, π_2), the standard simulation lemma (applied to the augmented Markov game) gives:

$$|W_{1,i}^{\pi_1, \pi_2}(s_1; \hat{M}) - W_{1,i}^{\pi_1, \pi_2}(s_1; M)| \leq H \cdot \epsilon_K + H \cdot \max_{s,t} \|\hat{\rho}_{2,t}(\cdot|s) - \rho_{2,t}(\cdot|s)\|_1,$$

where $\hat{M}$ denotes the estimated model and M the true model. Both error terms vanish a.s. by Step 1. Let $\delta_K := H \cdot \epsilon_K + H \cdot \max_{s,t} \|\hat{\rho}_{2,t} - \rho_{2,t}\|_1 \to 0$ a.s.

Step 3: Exploitability bound. Let $\hat{\boldsymbol{\pi}}_K = (\hat{\pi}_1, \hat{\pi}_2)$ be a CNE of the estimated model $\hat{M}$. By definition, for any counterfactual deviation π_1':

$$W_{1,1}^{(\pi_1', \hat{\pi}_2)}(s_1; \hat{M}) \leq W_{1,1}^{\hat{\boldsymbol{\pi}}_K}(s_1; \hat{M}).$$

Applying the simulation bound from Step 2 to both the left-hand side (replacing $\hat{M}$ with M) and using the triangle inequality:

$$W_{1,1}^{(\pi_1', \hat{\pi}_2)}(s_1; M) \leq W_{1,1}^{(\pi_1', \hat{\pi}_2)}(s_1; \hat{M}) + \delta_K \tag{45}$$

$$\leq W_{1,1}^{\hat{\boldsymbol{\pi}}_K}(s_1; \hat{M}) + \delta_K \tag{46}$$

$$\leq W_{1,1}^{\hat{\boldsymbol{\pi}}_K}(s_1; M) + 2\delta_K. \tag{47}$$

Taking the supremum over π_1' and applying the same argument symmetrically for agent 2:

$$\max_{i \in \{1,2\}} \sup_{\pi_i' \in \Pi_i^{\mathrm{CTF}}} \left(W_{1,i}^{(\pi_i', \hat{\boldsymbol{\pi}}_{K,-i})}(s_1; M) - W_{1,i}^{\hat{\boldsymbol{\pi}}_K}(s_1; M) \right) \leq 2\delta_K \to 0 \quad \text{a.s.}$$

This establishes that the exploitability vanishes almost surely. Any limit point of $\{\hat{\boldsymbol{\pi}}_K\}$ (which exists by compactness of the policy space) satisfies zero exploitability, hence is a CNE. □

D.3 Discussion of Identifiability

The linear independence conditions (Assumptions D.1–D.2) are the sequential analogue of the identifiability conditions used in Maiti et al. [2025] for single-stage causal games. Several remarks are in order.

When transitions do not depend on x_2'. If the state transition satisfies $P_t(s' \mid s, a_1, x_2, x_2') = P_t(s' \mid s, a_1, x_2)$ for all x_2'—i.e., the opponent's natural intuition affects only rewards, not dynamics—then Assumption D.2 is unnecessary. The transition is directly estimable from the exploration data without any identification step, and only Assumption D.1 is needed.

Relationship to finite mixture identification. The identification problem at each (s, a_1, x_2, t) is an instance of finite mixture deconvolution with known mixing weights. Classical results [Yakowitz and Spragins, 1968] establish that linear independence of the component distributions is both necessary and sufficient for identifiability of finite mixtures. The known-weights setting is strictly easier than the unknown-weights case, which would require Kruskal's condition or tensor decomposition methods.

Adversarial exploration. Algorithm 1 assumes that both agents cooperate during exploration by playing uniformly. If the opponent could deviate from the agreed exploration policy, the mixing weights become unknown, and identification requires jointly estimating the weights and components. This harder problem would require stronger structural conditions (e.g., Kruskal's condition for tensor decomposition) and is left to future work.

E Experiments

This appendix expands on the windy MPE experiments of Sec. 4. Sec. E.1 details the environment, the wind machinery, and the configurations compared; Sec. E.2 specifies the deep MARL algorithm (MADDPG with counterfactual inputs); Sec. E.3 reports the full quantitative results, including the wind-conditional breakdown and a discussion of variance and on-policy versus off-policy effects; Sec. E.4 lists the run identifiers and reproducibility information.

E.1 Environment: Windy Multi-Particle Environment

Motivation. Wind, current, and similar physical disturbances are first-order concerns for delivery drones, autonomous boats, mobile robots in agriculture, and search-and-rescue fleets, where gusts simultaneously push the objects of interest and fog the sensors that track them. The windy MPE captures this confound in a controlled MARL benchmark: the same disturbance that occludes the agents also moves the things they care about, so adversarial noise on static landmarks is not a faithful surrogate.

We extend two canonical MPE tasks with a wind perturbation that intermittently blacks out agent observations and applies a global force to all targets:

- **windy_simple_spread** (cooperative): $N = 3$ agents and $N = 3$ targets. The per-agent reward is the global negative sum of minimum agent-to-target distances minus per-agent collision penalties (unchanged from the canonical `simple_spread`). Episode length $H = 50$.
- **windy_simple_adversary** (mixed cooperative–competitive): one adversary, two good agents, two targets. One target is the secret goal goal_a. The good agents' reward rewards proximity of the closest good agent to the goal and distance of the adversary from the goal; the adversary's reward is the negative distance to the goal. Episode length $H = 75$.

The targets being movable is what makes the wind disturbance physically coherent: the same gust that blacks out the agents' sensors also pushes the very objects the agents are trying to track.

Wind machinery. A per-tick state machine drives the wind: (i) with probability p_{start} per step (when calm), a gust starts; the gust direction is sampled uniformly from $\{\text{left}, \text{right}, \text{up}, \text{down}\}$; (ii) a gust persists for T_{wind} steps, during which each target's velocity receives an additive force of magnitude w_{sens} in the gust direction (positions are integrated under the same damped Euler step

Table S1: 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%

used for agents and clamped to $[-1, +1]$), and each agent's *visible* observation slots (positions and velocities of self and others, relative target positions) are replaced with the sentinel value -1—a perceptual blackout while the wind is active; (iii) after T_{wind} steps the gust ends and observations return to normal. There is no minimum cooldown between gusts.

Counterfactual signal. While visible perception is blacked out, each counterfactual agent receives an additional slot appended to its observation vector. The slot encodes the greedy discrete action toward that agent's target—nearest target for spread, the goal target goal_a for adversary. During calm periods the slot is zero. The slot is per-agent and is computed greedily from the *true* underlying state, so it provides privileged information that survives the perceptual blackout. Operationally, the slot is the L_3 natural action $X'_{i,t}$ of Def. 2.2: the action agent i *would* take under its instinctive policy given the unperturbed state.

For discrete-action environments the env emits the slot as a single scalar in $\{0, 1, 2, 3, 4\}$ (the action index). The trainer one-hot expands this to a 5-dimensional vector before feeding it to the actor and the centralised critic, eliminating the impedance mismatch between integer-valued action indices (which have no ordinal meaning) and a single continuous network input.

Configurations compared. For each environment we run two configurations sharing all wind and environment hyperparameters:

- **Non-causal** (`scenario=none -no-intuition`): the env emits no L_3 slot at all. Agents see the perceptual blackout but no privileged signal during wind. This corresponds to the standard interventional (L_2) policy class.
- **Counterfactual** (L_3 policies that condition on the natural-action slot):
  - For spread: every agent (`scenario=all`) receives the slot pointing at its nearest target.
  - For adversary: only the two good agents receive the slot (`scenario=mixed`, `-intuition-mask "adversary_0=0,agent_0=1,agent_1=1"`), pointing at goal_a. The adversary remains non-causal during wind, so the L_3 signal asymmetrically advantages cooperators over the antagonist.

E.2 Algorithm: MADDPG with Counterfactual Inputs

We use Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [Lowe et al., 2017] with centralised training and decentralised execution: a per-agent deterministic actor receives only that agent's local observation (with the natural-action slot appended where applicable), while a per-agent centralised critic receives the global state and the joint action of all agents. This is following the Causal MADDPG schema of Sec. 3.3 (pseudocode in Algorithm 3).

Architecture. All networks are 64×64 MLPs with ReLU activations and orthogonal weight initialisation (gain $\sqrt{2}$ for hidden, 0.01 for the actor head, 1.0 for the critic head). The per-agent actor π_i takes the local observation (with the L_3 slot expanded if applicable) and outputs logits over 5 discrete actions; the per-agent critic Q_i takes the global state plus the concatenation of all agents' one-hot actions and outputs a scalar. Target networks π'_i and Q'_i mirror the online networks. The

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}.

for env step $= 1, 2, \ldots$ **do**
  Observe s_t; each agent i receives $x'_{i,t} \leftarrow f_{X_i}(s_t, U_{i,t})$ and local view $o_{i,t}$ with $x'_{i,t}$ appended.
  Each agent i samples $x_{i,t} \sim \mathrm{GS}_{\tau_{\mathrm{GS}}}\big(\pi_{\theta_i}(o_{i,t})\big)$ (straight-through one-hot).
  Execute $\mathbf{x}_t$; observe $\{r_{i,t}\}$, s_{t+1}; store $(s_t, \mathbf{x}'_t, \mathbf{o}_t, \mathbf{x}_t, \mathbf{r}_t, s_{t+1}, \mathbf{x}'_{t+1}, \mathbf{o}_{t+1})$ in $\mathcal{D}$.
  if step $\bmod T_{\mathrm{train}} = 0$ **and** $|\mathcal{D}| \geq T_{\mathrm{warm}}$ **then**
    Sample minibatch from $\mathcal{D}$.
    for each agent i **do**
      $y_i \leftarrow r_i + \gamma\,(1-d)\,Q_{\phi'_i}\big(s', \mathbf{x}'',\, \pi_{\theta'_1}(o'_1), \ldots, \pi_{\theta'_n}(o'_n)\big)$ *// target*
      $\phi_i \leftarrow \phi_i - \eta_Q\,\nabla_{\phi_i}\big(Q_{\phi_i}(s, \mathbf{x}', \mathbf{x}) - y_i\big)^2$ *// critic step*
      $\theta_i \leftarrow \theta_i + \eta_\pi\,\nabla_{\theta_i}\,Q_{\phi_i}\big(s, \mathbf{x}', x_1, \ldots, \pi_i^{\mathrm{ST}}(o_i), \ldots, x_n\big)$ *// 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

global state is the concatenation of all agents' local observations, each with the L_3 slot expanded to one-hot before concatenation.

Discrete actions via Gumbel–Softmax. To allow gradients to flow from the critic back through the actor's action sampling, the actor uses a Gumbel–Softmax sample with the straight-through estimator: the forward pass is the hard one-hot sample (passed to the environment as an integer action index), and the backward pass uses the soft Gumbel–Softmax gradient with temperature $\tau_{\mathrm{GS}} = 1.0$. Target actors use deterministic argmax (no Gumbel noise).

Training procedure. Each environment step: per-agent actor produces an action; the joint action is applied to the env; the transition (o_i, x_i, r_i, o'_i) for each agent, together with the global state and next state, is pushed to a shared replay buffer. Periodic update: every `train_every` $= 100$ environment steps after a warmup of $10{,}000$ environment steps, a batch of 1024 transitions is sampled and one gradient step per agent is performed. The critic loss is the squared TD error

$$\mathcal{L}^{(i)}_{\mathrm{critic}} \;=\; \big(Q_i(s, x_1, \ldots, x_N) \;-\; \big(r_i + \gamma\,(1-d)\,Q'_i(s', \pi'_1(o'_1), \ldots, \pi'_N(o'_N))\big)\big)^2, \tag{48}$$

and the actor loss is the deterministic policy gradient surrogate

$$\mathcal{L}^{(i)}_{\mathrm{actor}} \;=\; -\,Q_i\big(s,\, x_1, \ldots, x_{i-1}, \pi_i^{\mathrm{ST}}(o_i), x_{i+1}, \ldots, x_N\big), \tag{49}$$

where only the agent-i action contributes a gradient to π_i; other agents' actions come from the buffer (detached), and $\pi_i^{\mathrm{ST}}(o_i)$ is the Gumbel–Softmax straight-through sample of the current actor. Target networks are updated by Polyak averaging $\theta'_i \leftarrow \tau\,\theta_i + (1-\tau)\,\theta'_i$ with $\tau = 0.005$. Gradient clipping at L_2 norm 0.5 is applied to both actor and critic.

Implementation details. The full set of optimiser and training hyperparameters is given in Tbl. S2. Time-limit truncations are *not* treated as terminal states for the bootstrap: in MPE the next state after `max_cycles` is well-defined, so masking the bootstrap at every truncation systematically biased Q low and prevented learning. Done flags propagate into the bootstrap target only on real terminations (which do not occur in either MPE task here). The trainer collects experience from $n_{\mathrm{envs}} = 12$ `ParallelEnv` subprocess workers stepping synchronously each tick; one shared replay buffer aggregates transitions across workers. Per tick, each agent's actor processes a single batched forward pass of shape $(n_{\mathrm{envs}}, \dim\, o)$, and observation preprocessing (L_3-slot one-hot expansion plus global state concatenation) is performed in numpy to avoid small-tensor PyTorch dispatch overhead.

E.3 Full Results

All runs use a single seed (42) and approximately $50{,}000$ training episodes per run. Final deterministic evaluation uses 300 fresh episodes per run with the seed schedule offset to $999{,}000 +$ episode index,

Table S2: MADDPG training hyperparameters (shared across environments and configurations).

hyperparameter	value
optimiser	Adam (default β)
actor lr / critic lr	$7e{-4}$ / $7e{-4}$
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}	12
total episodes per run	$\sim$50,000

Table S3: Aggregate evaluation reward over 300 deterministic episodes per run. Reward is the per-agent average across cooperative agents (3 on spread, 2 good agents on adversary). The standard error of the difference is 0.64 (spread) and 1.53 (adversary). Both differences are highly significant ($z \gg 1.96$).

environment	configuration	reward	std	difference	z
spread (cooperative)	counterfactual (all)	-24.00	6.91	—	—
	non-causal	-31.93	8.72	$+\mathbf{7.93}$	12.34
adversary (good-agents avg)	counterfactual (good only)	$+26.88$	23.93	—	—
	non-causal	$+5.86$	11.53	$+\mathbf{21.01}$	13.70
adversary ($\mathrm{adversary}_0$)	counterfactual (good only)	-57.23	28.74	—	—
	non-causal	-38.28	11.20	—	—

yielding the *same* wind realisations across compared configurations and supporting paired comparisons by wind bin.

Aggregate reward. Tbl. S3 reports the deterministic evaluation reward over 300 episodes for each (environment, configuration) pair. Counterfactual agents improve per-agent reward by $+7.93$ on spread and by $+21.01$ on the adversary task's good agents, with z-scores of 12.34 and 13.70 respectively. The asymmetric design of the adversary configuration is also reflected in the adversary's own reward, which drops from -38.28 to -57.23 when the good agents are counterfactual—the good agents leverage their privileged signal directly against the adversary.

Wind-conditional reward. Because the eval seed sequence yields the same wind distribution across compared runs, we can perform a within-bin paired comparison by stratifying episodes by the number of wind events. Tbl. S4 reports the per-agent reward within each wind-event bin; Fig. 3c in the main text plots the same data with one-standard-error bars.

Analysis. Four observations emerge from the wind-conditional view.

(1) Counterfactual agents win in both environments. The aggregate per-agent reward gap is $+7.93$ on spread and $+21.01$ on adversary's good agents ($z = 12.34$ and 13.70), and the within-bin paired comparisons confirm the gap is not a noise artefact.

(2) The benefit scales with wind activity. On spread the gap grows monotonically from $+7.06$ (no wind) to $+8.49$ (≥ 3 events)—a modest but consistent increase. On adversary the gap is much more wind-sensitive: from $+3.09$ to $+29.39$, a factor of $\sim 9.5\times$. The L_3 signal is most valuable precisely in the regimes the perturbation creates: when perception is degraded, the privileged natural-action signal carries the information lost from the observation.

Table S4: Wind-conditional per-agent reward. Bins correspond to the number of wind events that occurred during the evaluation episode. The gap is computed as counterfactual − non-causal. Within-bin comparisons are paired, since matched seeds produce identical wind realisations across configurations.

spread (cooperative coverage)				
wind events / episode	n	counterfactual (all)	non-causal	gap
0	29	-19.18	-26.24	$+7.06$
1	79	-20.32	-28.06	$+7.74$
2	93	-24.55	-32.31	$+7.76$
≥ 3	99	-27.83	-36.32	$\mathbf{+8.49}$
adversary (good-agents avg)				
wind events / episode	n	counterfactual (good only)	non-causal	gap
0	9	$+11.81$	$+8.72$	$+3.09$
1	24	$+11.82$	$+6.80$	$+5.02$
2	92	$+16.66$	$+5.64$	$+11.02$
≥ 3	175	$+35.09$	$+5.70$	$\mathbf{+29.39}$

(3) The asymmetric scaling on adversary reflects an asymmetric problem. In the cooperative spread task the counterfactual agent's strategy is qualitatively the same with or without wind: head toward the nearest target. The L_3 slot recovers the lost perception during wind but does not change the optimal policy. In the adversary task the L_3 signal is asymmetrically given to the good agents, and during wind the adversary loses both perception of agent positions and any indirect signal about the goal. The good agents, who alone retain the goal pointer, can exploit the adversary's blindness in episodes with sustained wind.

(4) The privileged-information failure mode observed for MAPPO does not appear here. In an earlier MAPPO experiment at the same wind probability and a higher learning rate ($\mathrm{lr} = 2\mathrm{e}{-3}$, no entropy-bonus modification), the adversary learned over time to exploit the predictable L_3-driven policy of the good agents, causing a *reversal* in which the non-causal configuration outperformed the counterfactual configuration after sufficient training. With MADDPG we observe a clean win for the counterfactual configuration at every wind level. Two factors plausibly contribute: (i) the replay buffer keeps stale transitions in the training distribution, slowing the rate at which the adversary can co-adapt to the good agents' current policy; and (ii) the centralised critic supplies a smoother gradient signal that does not depend on tight on-policy estimates of the opponent's distribution. The combination produces good-agent policies that use the natural-action signal without becoming brittle in self-play. This is consistent with the remark in Sec. 3.3 that the counterfactual-input augmentation is orthogonal to the choice of underlying deep MARL algorithm; off-policy methods with experience replay appear better suited to the resulting non-stationarity than on-policy actor–critic methods at the learning rates used here.

Variance in adversary increases with wind. The standard deviation of good-agents reward in the counterfactual configuration (23.93) is more than double that of the non-causal configuration (11.53). Episodes with heavy wind (reward $\sim +35$) and episodes with no wind (reward $\sim +12$) draw the distribution much wider; quantifying performance with the conditional breakdown above is therefore more informative than a single aggregate mean.

E.4 Reproducibility

All runs use seed 42. Total environment steps per run are $n_{\mathrm{envs}} \times \mathrm{episodes} \times H \approx 30\mathrm{M}$ (spread) and $\approx 45\mathrm{M}$ (adversary), running in approximately 1.5–2 hours of wall time on 12 CPU cores. The eval seed schedule is $999{,}000 +$ episode_index for both compared runs, ensuring matched wind distributions. Run identifiers are `spread_blind_p05_maddpg_s42`, `spread_intuition_p05_maddpg_s42`, `adversary_blind_p05_maddpg_s42`, and `adversary_mixed_good_p05_maddpg_s42`.

E.5 Iterated Causal Prisoner's Dilemma

This subsection expands on Sec. 4.2: the Causal IPD instance, the identification pipeline, and the evaluation protocol behind Fig. 3.

Causal IPD instance. We use the IPD of Example 1.1 in environment M_1 ($R_{1,t} = R_{2,t} = 1$, $P(U_{i,t} = 0) = 0.6$ independently across rounds and agents). Per-round payoffs follow Fig. 1a; cumulative reward sums per-round payoffs without discounting.

Exploration. Both agents play uniform-random executed actions for $K_{\text{exp}} = 100{,}000$ rounds. Agent 1 records its own natural intuition $x'_{1,t}$, the joint executed action $(x_{1,t}, x_{2,t})$, and the reward $r_{1,t}$; agent 2's natural intuition is never observed. Because the exploration policy is uniform, the marginal $\rho_2(\cdot) = P(X'_{2,t} = \cdot)$ depends only on the environment and is recoverable from the pooled reward distribution as in Sec. 3.2. The budget 100K is chosen so that each (a_1, x_2) identification cell — with $a_1 = (x'_1, x_1) \in \mathcal{A}_1 \times \mathcal{A}_1$ — receives $\sim$12.5K samples in expectation under uniform play, well above the moment-matching threshold imposed by the linear-independence margin.

Identification. The IPD has no state transitions beyond the round counter, so $\textsc{Ident}_P$ is a no-op. $\textsc{Ident}_r$ is implemented by moment matching: at each (a_1, x_2), the empirical reward distribution is a mixture indexed by $x'_2 \in \{0, 1\}$ with known weights $\hat{\rho}_2$, and the component reward means are recovered by solving the resulting linear system, which is well-posed under Asm. D.1. Marginals $\hat{\rho}_1, \hat{\rho}_2$ are estimated from pooled samples.

Exploitation. With the recovered payoffs, the exploitation phase computes a stage $\textsc{BayesNash}$ on the augmented payoff matrix indexed by (x'_1, x'_2), yielding a stationary L_3 policy $\hat{\pi}_1(\cdot \mid x'_1)$ for agent 1.

Evaluation. The learned policy is deployed for $T_{\text{eval}} = 10{,}000$ rounds against (i) tit-for-tat (TfT), which mirrors the last executed action, and (ii) always-defect (D). Both opponents are deterministic; trajectory stochasticity comes solely from agent 1's natural-action draws $X'_{1,t}$. Cumulative-reward trajectories are reported in Fig. 3.

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