TECHNICAL REPORT
R-125
First version: Jan, 2025
Last version: July, 2025

Counterfactual Rationality: A Causal Approach to Game Theory

Aurghya Maiti and **Prateek Jain** and **Elias Bareinboim**
Causal Artificial Intelligence Laboratory
Columbia University, USA
{aurghya,eb}@cs.columbia.edu

Abstract

The tension between rational and irrational behaviors in human decision-making has been acknowledged across a wide range of disciplines, from philosophy to psychology, neuroscience to behavioral economics. Models of multi-agent interactions, such as von Neumann and Morgenstern's expected utility theory and Nash's game theory, provide rigorous mathematical frameworks for how agents should behave when rationality is sought. However, the rationality assumption has been extensively challenged, as human decision-making is often irrational, influenced by biases, emotions, and uncertainty, which may even have a positive effect in certain cases. Behavioral economics, for example, attempts to explain such irrational behaviors, including Kahneman's dual-process theory and Thaler's nudging concept, and accounts for deviations from rationality. In this paper, we analyze this tension through a causal lens and develop a framework that accounts for rational and irrational decision-making, which we term *Causal Game Theory*. We then introduce a novel notion called counterfactual rationality, which allows agents to make choices leveraging their irrational tendencies. We extend the notion of Nash Equilibrium to counterfactual actions and Pearl Causal Hierarchy (PCH), and show that strategies following counterfactual rationality dominate strategies based on standard game theory. We further develop an algorithm to learn such strategies when not all information about other agents is available.

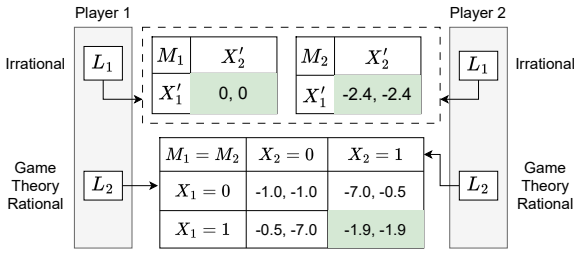
1 Introduction

Decision-making in multi-agent systems (MAS) is a critical problem with broad applications across disciplines such as economics, social sciences, political science, distributed systems, robotics, and more recently, in aligning AI systems with human preferences. At its core, such decision-making involves taking into account multiple agents – individuals, autonomous systems, or organizations – each with their own objectives, preferences, and constraints, to make coherent and coordinated decisions within complex, dynamic environments. The complexity of decision-making in MAS arises from the interplay of several factors, including uncertainty, inherent biases, conflicting objectives, and the limitations of the agents' computational and observational capabilities.

Von Neumann and Morgenstern [45] reformulated and popularized *expected utility theory* [35], laying the foundation for *rational* decision-making, where agents select actions to maximize their expected utility. Since then, Game Theory (GT) has become central to MAS, with models, such as Nash equilibrium [29], cooperative game theory [38], evolutionary game theory, and Bayesian games [16], offering tools to analyze scenarios where agents' choices impact one another. Although rational decisions are grounded in systematic analysis and objective reasoning, human choices are often influenced by cognitive biases, emotions, social factors, and various unobserved factors that lead to seemingly irrational outcomes. Sometimes, irrational or naive choices can even result in better

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

(a)



(b)

Figure 1: (a) Y_1, Y_2 as a function of U_1, U_2, X_1, X_2 (b) To be rational or not to be rational, that is the question.

outcomes than rational ones, a phenomenon known as *paradox of rationality* [17, 7, 5]. Behavioral economics seeks to model such deviations from rationality, with models such as loss aversion [21], anchoring [42], framing of choices [22], social preferences [11], and emotions [28]. Kahneman [20] also advanced and popularized *dual-process theory* [46, 39], which posits two cognitive systems: a fast automatic *System 1* and a slow deliberate *System 2*. While these approaches help explain aspects of irrational human decision-making, the broader question of *when* and *how* such unobserved biases can be strategically leveraged in MAS remains largely unexplored.

In this work, we make a significant step towards addressing this gap by proposing a framework rooted in causal modeling [31, 4]. Human decisions are often guided by causal structures [44, 40, 30], and actions can be viewed as interventions [13]. Building on these insights, we model the environment and the agent's decision-making process as an interplay between exogenous and endogenous factors, represented as a structural causal model (SCM). SCMs have been used successfully in the context of decision-making, both for single-step bandit problems [3, 48] and for multistep RL settings [27, 37], as surveyed in [2]. The advantage of such modeling is not only computational but more fundamental. Consider the example of *Greedy Casino* in Bareinboim et al. [3], where a randomized control trial (RCT) suggests that the expected payoff is higher than the realized payoff of players following their natural instincts (irrational behavior). One may naturally surmise that, given the superiority of the automated version based on RCTs, humans and their irrationality could be removed from the loop. However, players could enact a counterfactual randomization procedure that exploits their natural biases, which surprisingly led to payoffs exceeding those based on the RCT.

Building on these insights, we model MAS through a causal lens and show that existing game models may not capture similar fundamental features of the decision-making process. This framework models the interactions of agents within a system through the different layers of PCH [4]. As a consequence, an agent will have the capability to act rationally (following Nash's prescription), instinctively, or as some mixture of both. We introduce the notion of *counterfactual rationality* to formally determine when it is advantageous for agents to act irrationally and when it is better to avoid doing so. The next example illustrates why this task is nontrivial.

Example 1.1 (Causal Prisoner's Dilemma (CPD))**.** Two thieves are suspected of a crime, but due to insufficient evidence, they cannot be convicted outright. Now, they have a choice to make – either remain silent (cooperate, C) or betray the other (defect, D). We denote the choices by variables X_1 and X_2, and cooperation and defection by the values 0 and 1, respectively. The thieves' decisions are influenced by external circumstances, represented by variables U_1 and U_2, which capture factors such as the temperament of police officers, the competence of legal defense, new evidence or witnesses emerging, and even the disposition of the judge and the jury. Although these factors cannot be explicitly measured by the prisoners, they may subconsciously shape their decisions.

Each prisoner has a natural ability to assess their circumstances, denoted by R_1 and R_2. If prisoner i has an accurate reading of their situation ($R_i = 1$), they choose to cooperate ($X_i = 0$) if the circumstances are favorable ($U_i = 1$), and defect when they are adversarial ($U_i = 0$); conversely, if they have a poor reading of their situation ($R_i = 0$), they defect when circumstances are good, and cooperate when circumstances are bad. For prisoner i, their instinctive or natural choice is modeled as: $X_i \leftarrow f_X(R_i, U_i) = R_i \oplus U_i$, where $\oplus$ is the exclusive-or operator. We note that the variables U_1, U_2, R_1, R_2 and the function f_X are determined by nature and are unknown to the prisoners.

Now, we analyze two scenarios, M_1 and M_2. In M_1, the prisoners have a good reading of their situation ($R_1 = R_2 = 1$), while in M_2, they misjudge their circumstances ($R_1 = R_2 = 0$). In both cases, $P(U_i = 0) = 0.6$ for $i \in \{1, 2\}$. The outcome $\mathbf{Y} = (Y_1, Y_2)$ of their decisions is a function of U_1, U_2, X_1 and X_2 as shown in Fig. 1a. For example, when the situation is favorable for both the prisoners ($U_1 = 1, U_2 = 1$) and they cooperate ($X_1 = 0, X_2 = 0$), their payoff is $(0, 0)$. However, if circumstances are favorable for Prisoner 1 and not for Prisoner 2 ($U_1 = 1, U_2 = 0$), and Prisoner 1 defects while Prisoner 2 cooperates ($X_1 = 1, X_2 = 0$), their payoff is $(0, -8)$.

If both prisoners ignore their intuition and search for the optimal strategy, the situation corresponds to the classical Prisoner's Dilemma, where the payoff for the actions $X_1 = x_1, X_2 = x_2$ is given by:

$$\sum_{u_1, u_2, \mathbf{y}} \mathbf{Y} \cdot P(u_1, u_2) P(\mathbf{Y} \mid x_1, x_2, u_1, u_2) \tag{1}$$

Notably, both scenarios M_1 and M_2 lead to the same Prisoner's Dilemma (PD) game, as shown in the 2×2 payoff table at the bottom of Fig. 1b. However, if both prisoners rely on their natural instincts, their expected payoff is $(0, 0)$ in M_1 and approximately $(-2.4, -2.4)$ in M_2. This is illustrated in Fig. 1b, where X_1' and X_2' denote the players acting based on their natural intuition (shown in the top row). The situation presents a new dilemma – it is better to follow natural instincts and be irrational in M_1, whereas it is better to be rational and ignore intuition in M_2.

This example raises a fundamental question: when is it better to follow natural intuition and when is it better to override it and follow Nash's prescription? In this paper, we explore the tension between rational and instinctive behavior through a causal lens and derive from first principles how agents should deliberate and make decisions, thus addressing the so-called 'paradox of rationality' (see Appendix A). Specifically, we outline our technical contributions as follows:

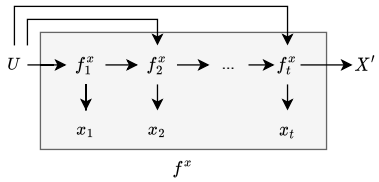
1. We formalize a class of games that combine rational and irrational behavior (Def. 2.10) and show that it strictly generalizes traditional Normal Form Games (Thm. 2.11).
2. We introduce a new family of *counterfactual strategies*, prove the existence of equilibrium (Thm. 3.5), and show that these strategies can outperform other strategies (Thm. 3.6).
3. We develop an algorithm `CTF-Nash-Learning` (Alg. 2) that learns the payoff matrix in the counterfactual action space and identifies equilibria, even when the actions of the other agents are not fully observed.

Preliminaries. In this section, we introduce the notations and definitions used throughout the paper. We use capital letters to denote random variables (X) and small letters to denote their values (x). $\mathcal{D}_X$ denotes the domain of X. $|\mathbf{S}|$ denotes the cardinality of the set $\mathbf{S}$. The basic framework of our model resides on Structural Causal Models [31]. An SCM M is a tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where $\mathbf{V}$ and $\mathbf{U}$ are sets of endogenous and exogenous variables, respectively. $\mathcal{F}$ is a set of functions f_V determining the value of $V \in \mathbf{V}$, that is, $V \leftarrow f_V(\mathbf{Pa}(V), \mathbf{U}_V)$, where $\mathbf{Pa}_V \subseteq \mathbf{V}$ and $\mathbf{U}_V \subseteq \mathbf{U}$. Naturally, M induces a distribution over the endogenous variables, $P(\mathbf{V})$, called *observational or L_1 distribution*. An intervention on a subset $\mathbf{X} \subseteq \mathbf{V}$, denoted by $do(\mathbf{x})$ is an operation where values of $\mathbf{X}$ are set to $\mathbf{x}$, replacing the functions $\{f_X : X \in \mathbf{X}\}$. For an SCM M, $M_{\mathbf{x}}$ denotes the model induced by the operation $do(\mathbf{x})$ and $P_{\mathbf{x}}(\mathbf{Y})$ or $P(\mathbf{Y}_{\mathbf{x}})$ denotes the probability of $\mathbf{Y}$ in $M_{\mathbf{x}}$. Such distributions are called *interventional or L_2 distributions*. For further details and discussions on counterfactual distributions, refer to Appendix A.1 and Bareinboim et al. [4, Sec.1.2]. Additional background and examples on decision-making in single-agent causal systems can be found in Bareinboim et al. [2] and Appendix A.5, along with comparisons to related work [15, 12] in Appendix A.

2 Causal Normal Form Games

In this section, we model the interaction of multiple agents in a system through the language of SCMs and PCH layers. Here, we generalize the concepts introduced in [2] to multi-agent settings. We first define a set of action nodes and reward signals for the agents in the system along with the SCM.

Definition 2.1 (Causal Multi-Agent System)**.** A Causal Multi-Agent System (CMAS) is a tuple $\langle M, N, \mathbf{X}, \mathbf{Y} \rangle$, where (i) $M : \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, \mathbb{P} \rangle$ is an SCM, (ii) N is the set of n agents, (iii) $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is a tuple of action nodes with disjoint $\mathbf{X}_i, \mathbf{X}_j \subset \mathbf{V}$ for $i, j \in [n]$, $i \neq j$, and (iv) $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ is the ordered set of reward signals, with $\mathbf{Y}_i \subseteq \mathbf{V} \setminus \mathbf{X}$ for all $i \in [n]$. □



(a)

P1 \ P2	$X_2 = X_2'$	$X_2 = 0$	$X_2 = 1$
$X_1 = X_1'$	$-2, 2$	$-2, -2$	$-2, -2$
$X_1 = 0$	$0, 0$	$-1.5, -1.5$	$-1.5, -1.5$
$X_1 = 1$	$0, 0$	$-1.5, -1.5$	$-1.5, -1.5$
$X_1 = 1 - X_1'$	$2, -2$	$-1, -1$	$-1, -1$

(b)

Figure 2: (a) Illustration of decision flow f_X (b) It is not always optimal to jump to L_3 policy

A CMAS is essentially an SCM with a set of action nodes $\mathbf{X}$, each controlled by one of the n agents. In addition, the system includes reward variables, $\mathbf{Y}$, representing the feedback each agent receives based on their actions and the underlying causal mechanism.

Example 2.2. Consider the CPD presented in Ex. 1.1. The SCM $\mathcal{M}$ corresponding to scenario M_2 is defined as: (i) $\mathbf{U} = \{U_1, U_2, R_1, R_2\}$, $\mathbf{V} = \{X_1, X_2, Y_1, Y_2\}$, (ii) $X_i = R_i \oplus U_i$ for $i \in \{1, 2\}$. Y_1, Y_2 as a function of U_1, U_2, X_1, X_2 are shown in Fig. 1a, and (iii) $P(U_i = 1) = 0.4, P(R_i = 0) = 1$ for $i \in \{1, 2\}$. The CMAS can now be defined as $\langle M = \mathcal{M}, N = \{1, 2\}, \mathbf{X} = (\{X_1\}, \{X_2\}), \mathbf{Y} = (\{Y_1\}, \{Y_2\})$. □

Now, we define different forms of actions that an agent may take in such a system. First, we define the different action and policy spaces and then explore how the action spaces are related.

Definition 2.3 (L_1 action)**.** Given a CMAS $\langle M, N, \mathbf{X}, \mathbf{Y}\rangle$, an L_1 action of an agent i is the one in which the value of their action variables $\mathbf{X}_i$ is determined by the natural mechanism $f_{\mathbf{X}_i} \in \mathcal{F}$. □

We will also call such actions *natural actions* and denote them by a_0. Note that, while performing a_0, an agent *does not know anything about the underlying SCM* nor do they deliberately change any mechanism of action variable in the system. The L_1 action space is thus $\mathcal{A}^1 = \{a_0\}$ and the L_1 policy space is also a singleton set $\Pi^1 = \{a_0\}$.

Example 2.4. Consider the CMAS presented in Ex. 2.2. The natural action is when the values of X_1 and X_2 are determined by their natural function, $X_1 = R_1 \oplus U_1, X_2 = R_2 \oplus U_2$ The expected payoff when both the agents are following their natural intuition is then given by

$$\sum_{u_1, u_2, x_1, x_2, \mathbf{y}} \mathbf{y} \cdot P(u_1, u_2) P(x_1 \mid u_1) P(x_2 \mid u_2) P(\mathbf{y} \mid u_1, u_2, x_1, x_2) \approx (-2.4, -2.4) \tag{2}$$

In traditional game-theoretic sense, an agent can intervene on the system via atomic interventions (setting action variables to fixed values based on context) [31], or soft interventions (sampling actions from a distribution) [9]. Next, we define L_2 actions and the associated policy space.

Definition 2.5 (L_2-action)**.** Given a CMAS $\langle M, N, \mathbf{X}, \mathbf{Y}\rangle$, L_2 action of an agent i is a hard intervention $do(\mathbf{x})$, where $\mathbf{x} \in \mathcal{D}_{\mathbf{X}_i}$. □

Hence, if an agent i performs $do(\mathbf{x}_i)$ in the SCM M, then the natural mechanism $f_{\mathbf{X}_i}$ is replaced by $\mathbf{X}_i \leftarrow \mathbf{x}_i$. The set of such L_2 actions is denoted by $\mathcal{A}^2$, and an L_2 policy is a distribution over $\mathcal{A}^2$.

Example 2.6. Consider the CMAS introduced in Ex. 2.2. L_2 action is when an agent performs an intervention, that is, setting their action variable to a particular value. If Player 1 is playing 0 and Player 2 is playing 1, then the assignment of the variables are given by $X_1 \leftarrow 0, X_2 \leftarrow 1$ and U_1, U_2, R_1, R_2 are sampled from $P(\mathbf{U})$ as in Ex. 2.2. Similarly, Y_1, Y_2 are determined by Fig. 1a. For instance, the expected payoff of the strategy $(do(X_1 = 0), do(X_2 = 1))$ will then be given by

$$\sum_{u_1, u_2, \mathbf{y}} \mathbf{y} \cdot P(u_1, u_2) P(\mathbf{y} \mid u_1, u_2, X_1 = 0, X_2 = 1) \approx (-7.0, -0.5) \tag{3}$$

It is also possible for one agent to perform an L_2 action and the other to perform an L_1 action. For instance, the payoff the strategy $(do(X_1 = 1), a_0)$ is given by

$$\sum_{u_1, u_2, x_2, \mathbf{y}} \mathbf{y} \cdot P(u_1, u_2) P(x_2 \mid u_2) P(\mathbf{y} \mid u_1, u_2, X_1 = 1, x_2) \approx (0, -8.9) \tag{4}$$

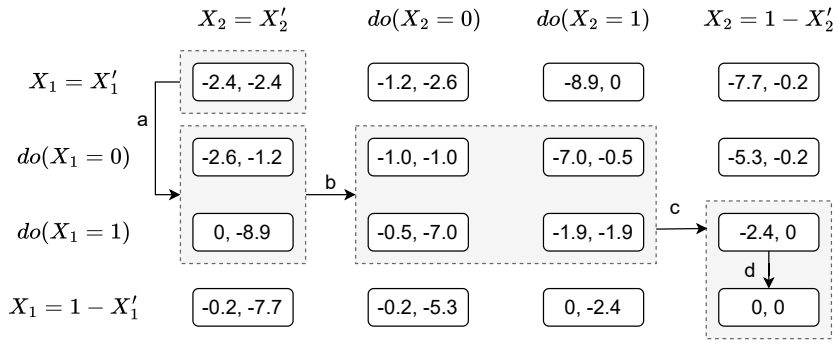


Figure 3: Change of Equilibrium with change of policies in Causal Prisoner's Dilemma.

In many cases, an agent can interact with the environment through PCH's Layer 3 [3, 4, 33], enabling counterfactual reasoning in their decision-making, and entering the realm of L_3 distribution. For example, in scenario M_2 of Ex. 1.1, following natural instinct led to a suboptimal outcome. However, if both agents had done the exact opposite of their instinctive choices, they could have achieved a payoff of $(0, 0)$. Now, we formally define L_3 actions.

Definition 2.7 (L_3-action space $\mathcal{A}^3$)**.** Given a CMAS $\langle N, M, \mathbf{X}, \mathbf{Y} \rangle$, an L_3 action for agent i is defined as a mapping $h : \mathcal{D}(\mathbf{X}_i) \to \mathcal{D}(\mathbf{X}_i)$ from intuition to action. □

When an agent takes an L_3 action, they first note their natural instinct $\mathbf{X}_i' \leftarrow f_{\mathbf{X}_i}(\mathbf{U}_i)$ and then executes $\mathbf{X}_i \leftarrow h_i(\mathbf{X}_i')$, where $\mathbf{U}_i$ is the set of unobserved parents of $\mathbf{X}_i$. If $h(x) = x$, it corresponds to the L_1 action, and if $h(x)$ is constant for all x, it is the L_2 action. Bearing this in mind, we will often denote a_0 as $\mathbf{X} = \mathbf{X}'$, where $\mathbf{X}$ is the action variable and $\mathbf{X}'$ is the intuition.

Bareinboim et al. [3] introduces a novel form of randomization to interact through the Layer 3 of PCH – interrupt any reasoning agent just before they execute their choice, treat this choice as their intention, and then act. This procedure involves subtle issues, and we refer readers to Sec. 7 in Bareinboim et al. [2] for a more detailed discussion. The agent may consider various options during the deliberation process, but only the final choice matters. For example, an agent may initially choose $X' = x_1$, then reconsider and change it to $X' = x_2$ and may continue doing so, until at time step t, it chooses $X' = x_t$ and decides to execute it. This final decision defines the agent's instinct irrespective of the path taken to reach it (see Fig. 2a). The same reference also proposed `Ctf-RCT`, where an intended action is perceived first, but instead of executing it directly, the final action is chosen uniformly at random from the entire action space. Now, we look at how to compute the payoff under L_3 action.

Example 2.8. Consider the CMAS in Ex. 2.2. An L_3 action would allow the agent to choose an action based on their natural intuition. Let g_1 and g_2 be two functions from $\{0, 1\}$ to $\{0, 1\}$. If the players are playing g_1 and g_2, respectively, then the variables are given by $X_i' \leftarrow R_i \oplus U_i, X_i \leftarrow g_i(X_i')$ for $i \in \{1, 2\}$. The variables U_1, U_2, R_1, R_2 are sampled from $P(\mathbf{U})$, and Y_1, Y_2 are determined by Fig. 1a. For example, if $g_1(x) = 1 - x$ and $g_2(x) = 1 - x$, then the expected payoffs are given by

$$\sum_{u_1, u_2, x_1, x_2, \mathbf{y}} \mathbf{y} \cdot P(u_1, u_2) P(x_1, x_2 \mid u_1, u_2) P(\mathbf{y} \mid u_1, u_2, g_1(x_1), g_2(x_2)) = (0, 0) \tag{5}$$

The payoffs for the various combinations of actions in Ex. 2.2 are shown in Fig. 3. Once the action spaces are defined, the policy space can be defined as a distribution over the action space. Let $\Delta(A)$ denote the set of distributions over the set of actions A. Then L_2 policy space $\Pi^2 = \Delta(\mathcal{A}^2)$ and L_3 policy space $\Pi^3 = \Delta(\mathcal{A}^3)$. Next, we define the notion of reward.

Definition 2.9 (Reward Function)**.** A reward function $\mathcal{R}_i : \mathcal{D}(\mathbf{Y}_i) \to \mathbb{R}$ of an agent i is a function from outcome $\mathbf{Y}_i$ to real numbers. □

In Ex. 1.1, we assume that the reward function is identity, that is, $\mathcal{R}_i(Y_i) = Y_i$ for $i \in \{1, 2\}$. Now that we have all the tools, we are ready to define Normal Form Games in proper causal language.

Definition 2.10 (Causal Normal Form Game)**.** A tuple $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$ is a Causal Normal Form Game (CNFG), where (i) $\mathbb{M}$ is a CMAS $\langle M, N, \mathbf{X}, \mathbf{Y} \rangle$, (ii) $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_n)$ is the set of policies for the n agents, $\mathcal{A}_i \subseteq \{\mathcal{A}^1, \mathcal{A}^2, \mathcal{A}^3\}$, and (iii) $\mathcal{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_n)$ is the set of reward functions. □

A CNFG is thus a CMAS, along with the policy space of the n agents and their reward functions. Now we will formally state the result generalizing our observation from CPD (Ex. 1.1).

Theorem 2.11. *Given a game in normal form, there exist two CNFGs $\mathcal{C}_1$ and $\mathcal{C}_2$ with equilibrium payoffs μ_1 and μ_2 under the action space $\mathcal{A}^1 \cup \mathcal{A}^2$, and a Nash Equilibrium (NE) payoff μ_{NE}, such that $\mu_2 < \mu_{NE} < \mu_1$ where $<$ denotes Pareto domination.* □

The theorem implies some important observations. CNFGs strictly generalize Normal Form Games (NFGs), capturing aspects such as instinctive behaviors and counterfactual policies that NFGs cannot naturally express. Although one might argue that CNFGs can be flattened into an equivalent NFG (Fig. 3), similar to Extensive Form or Bayesian Games, we claim causal modeling is not only advantageous but necessary: (i) Constructing the full payoff matrix requires an SCM, since actions are not arbitrary and defined only within that causal structure; (ii) NFGs do not clarify how actions are executed or whether agents are even capable of executing them; SCMs provide a concrete notion of agency; (iii) our solution concept presented in Sec. 3 relies on the hierarchical structure of the action spaces; (iv) finally, NFGs cannot capture the structure between intuitions and executed actions. In many cases, agents can only observe executed actions, and for computing equilibria, exploiting the structure becomes a necessity (Alg. 2). More details are provided in Appendix E.

3 Causal Nash Equilibrium

In this section, we introduce counterfactual rationality and establish the Causal Nash Equilibrium (CNE) for a CNFG. Allowing agents to transition between layers of the PCH leads to a two-step decision process. First, the agent determines which layers to operate in – instinct-based (L_1), classical rationality (L_2), or counterfactual reasoning (L_3). Second, the agent must decide which action to take within the chosen layer. We refer to this two-step process as a *causal strategy*. An agent is counterfactually rational if it seeks to maximize its expected payoff using causal strategies, given that other agents are also counterfactually rational. Next, we analyze how equilibrium outcomes change when agents move to higher layers of the PCH.

Example 3.1 (Equilibria in CPD)**.** Consider Ex. 1.1 (M_2) where we analyze how the payoffs and equilibria evolve as agents move across the layers of the PCH, from instinct-based L_1 policies to counterfactual-based, L_3 policies. Fig. 3 shows the payoff of the prisoners in this larger action space. If both prisoners follow their natural choices, playing L_1, their payoffs are $(-2.4, -2.4)$.

Now, suppose prisoner 1 starts thinking rationally, ignoring their natural instincts, which results in transition (a) in Fig. 3. Prisoner 1 eventually defects, meaning they play the action $do(X_1 = 1)$, while prisoner 2 still follows their instinct, $X_2' = X_2$. As a result, the payoffs become $(0, -8.9)$, where prisoner 1 benefits while prisoner 2 suffers. Eventually, prisoner 2 also learns to think rationally, leading to transition (b). In this case, both prisoners enter the realm of Standard Game Theory (SGT), each choosing to defect, playing the actions $(do(X_1 = 1), do(X_2 = 1))$. This results in NE with payoffs of $(-1.9, -1.9)$. A few observations are worth making at this point. First, the scope of SGT is highlighted in the four central cells of Fig. 3. Second, as noted earlier, the equilibrium in SGT is worse than when both agents act irrationally (L_1). The SGT analysis stops at this point, but our new framework suggests that strategic thinking may continue.

Over time, prisoner 2 introspects and contemplates counterfactual decisions, as highlighted in transition (c). They realize that their natural instincts provide insights that can be leveraged, and they should choose to act opposite to their natural choices, $X_1 = 1 - X_1'$. This yields payoffs of $(-2.4, 0)$, improving their baseline and hurting prisoner 1. Eventually, prisoner 1 also reaches L_3, leading to transition (d). Both players, now operating under Causal Game Theory (CGT), settle on actions against their natural instincts, $X_1 = 1 - X_1', X_2 = 1 - X_2'$, achieving payoffs of $(0, 0)$. This is the final state, where no unilateral deviation can increase payoffs. □

The game in this example reflects an increasingly refined form of human rationality, tracing its evolution from primitive instincts based on raw intuition (L_1) to a notion of rationality based on game theory, where the intuition is ignored (L_2), and going to advanced strategic thinking leveraging both rational and irrational aspects of human cognition (L_3).

A natural question that arises from this discussion is if it is always better to consider the full payoff table, since it provides the largest action space. To answer this, consider the example shown in Fig. 2b. The full game specification is given in Appendix D. If Player 1's action space is limited to L_1 and L_2,

P1 \ P2	$\mathcal{A}^1$	$\mathcal{A}^2$	$\mathcal{A}^1 \cup \mathcal{A}^2$
$\mathcal{A}^1$	$-2, 2$	$-2, -2$	$-2, 2$
$\mathcal{A}^2$	$0, 0$	$-1.5, -1.5$	$0, 0$
$\mathcal{A}^1 \cup \mathcal{A}^2$	$0, 0$	$-1.5, -1.5$	$0, 0$
$\mathcal{A}^3$	$2, -2$	$-1, -1$	$-1, -1$

(a)

P1 \ P2	$\mathcal{A}^1$	$\mathcal{A}^2$	$\mathcal{A}^1 \cup \mathcal{A}^2$	$\mathcal{A}^3$
$\mathcal{A}^1$	$-2.4, -2.4$	$-8.9, 0$	$-8.9, 0$	$-8.9, 0$
$\mathcal{A}^2$	$0, -8.9$	$-1.9, -1.9$	$-1.9, -1.9$	$-2.4, 0$
$\mathcal{A}^1 \cup \mathcal{A}^2$	$0, -8.9$	$-1.9, -1.9$	$-1.9, -1.9$	$-2.4, 0$
$\mathcal{A}^3$	$0, -8.9$	$0, -2.4$	$0, -2.4$	$0, 0$

(b)

Figure 4: Layer selection game for (a) example in Fig. 2b, and (b) Causal Prisoner's Dilemma.

then the equilibrium payoff is $(0, 0)$ (marked in blue). However, if the action space L_3 is considered, the last row in the table is also considered (gray), and the equilibrium payoff decreases to $(-1, -1)$. Hence, regardless of what the other player does, Player 1's mere consideration of a larger action space harms them. Broadly, deciding which action space to follow is non-trivial. Next, we define a projection of a CNFG, where action spaces are restricted to specific layers of the PCH.

Definition 3.2 (PCH Projection)**.** Given a CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$, the PCH projection of Γ, denoted by $\Gamma(A_1, \ldots, A_n)$, is the subgame of Γ where the action space of agent i is constrained to a subset $\mathcal{A}_i \supseteq A_i \in \{\mathcal{A}_i^1, \mathcal{A}_i^2, \mathcal{A}_i^1 \cup \mathcal{A}_i^2, \mathcal{A}_i^3\}$. □

This projection captures how a game evolves when agents operate within a restricted subset of available strategies corresponding to different levels of reasoning within the PCH. The problem now, is to find a projection from where agents have no incentive to unilaterally deviate to a different layer of the PCH. To address this, we introduce a strategic layer selection game, a meta-game, where agents choose which layer of PCH to operate at.

Definition 3.3 (Layer Selection Game)**.** Given a CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$, its Layer Selection Game L_Γ is the NFG with (i) the same set of agents N, (ii) action space $A = A_1 \times \ldots, A_n$, where $\mathcal{A}_i \supseteq A_i \in \{\mathcal{A}_i^1, \mathcal{A}_i^2, \mathcal{A}_i^1 \cup \mathcal{A}_i^2, \mathcal{A}_i^3\}$ and (iii) utility $u(A) = \texttt{NE}(\Gamma(A_1, \ldots, A_n))$ where $\texttt{NE}(\Gamma(A_1, \ldots A_n))$ is a Nash Equilibrium payoff of the CNFG Γ when actions spaces are restricted to $A_1, \ldots, A_n$. □

This metagame represents a higher-level decision process, where each cell in the payoff matrix corresponds to a PCH projection of Γ, and its equilibrium will determine the layer of reasoning in which the agents should operate. We will assume that such counterfactual rationality is common knowledge, that both players are aware that the other player can forget a part of their actions space and choose the PCH layers in which they operate. Let s_i^* be the NE strategy for player i in the layer selection game. Let $\text{supp}(s_i^*)$ denote the support of s_i^* – the set of action spaces with non-zero probability in s_i^*. In particular, if $\mathcal{A}_i^j \notin \text{supp}(s_i^*)$, then the agent can ignore, or "forget" about this action space, and instead play a PCH projection of Γ that excludes $\mathcal{A}_i^j$. For instance, in Fig. 2b, if Player 1 is able to forget that it can play L_3, the payoff for the agent is $(0, 0)$, which is higher than the payoff that with playing L_3, $(-1, -1)$.

In practice, agents can limit their reasoning layers by restricting their capabilities: (i) at L_1, agents act instinctively without requiring sampling mechanisms, (ii) at L_2, agents may need access to randomization (e.g., coin flips) for mixed strategies, and (iii) At L_3, agents must introspect, observe their intuition, and then decide how to act based on it, through more sophisticated procedures, such as ctf-randomization. Refusing to observe intuition renders L_3 inaccessible. One key observation is that forgetting part of the action space may not always be a good idea. For example, consider the simple prisoner's dilemma. If the agents choose to forget defect D and just play with the action space cooperate $\{C\}$, they will get a payoff $(-1, -1)$. However, one agent may start using the action space $\{C, D\}$ and then choose to defect, obtaining a payoff of -0.5 while the other agent gets -7.0. Thus it is not in the agent's interest to forget about defecting (see Appendix D).

Definition 3.4 (Causal Nash Equilibrium, or CNE)**.** Let Γ be a CNFG and L_Γ be its corresponding layer selection game with NE strategy s^*. A strategy profile π^* is called CNE if π^* is the Nash Equilibrium of $\Gamma(A^*)$, where $A^* = A_1 \times \ldots \times A_n$, and $A_i = \cup_{\mathcal{A} \in \text{supp}(s_i^*)} \mathcal{A}$. □

Theorem 3.5 (Existence of CNE)**.** *For any CNFG, CNE always exists.* □

If playing L_2 is a pure strategy NE of the layer selection game L_Γ, then the CNE of Γ in CGT and the NE of the normal form game induced by Γ coincide. Note that it is possible for a CNFG to have

Algorithm 1 `Find-CNE`

1: **Input:** PCH projections of CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$ **Output:** CNE strategies π^*
2: Construct the Layer Selection Game, L_Γ: For all $A = A_1 \times \ldots \times A_n$, such that $\mathcal{A}_i \supseteq A_i \in \{\mathcal{A}^1, \mathcal{A}^3, \mathcal{A}^1 \cup \mathcal{A}^2, \mathcal{A}^3\}$, $u(A) \in \texttt{NE}(\Gamma(A_1, \ldots, A_n))$
3: Let s^* be the NE strategy of L_Γ and $A^* = A_1^* \times \ldots A_n^*$, where $A_i^* = \bigcup_{\mathcal{A} \in \text{supp}(s_i^*)} \mathcal{A}$
4: **Return:** NE strategies of $\Gamma(A^*)$

Algorithm 2 `Ctf-Nash-Learning`

1: **Input:** Dataset from `Ctf-RCT`: $(x_1', x_1, x_2, \mathbf{y})$
2: **Output:** Causal Nash Equilibrium strategy f^*
3: For each (x_1', x_1, x_2), estimate the mean and weights of the distributions' mixture from the samples (y_1, y_2). Let the distribution means be $R_1(x_1', x_1, x_2), \ldots, R_k(x_1', x_1, x_2)$ with corresponding weights $p_1(x_1', x_1, x_2), \ldots, p_k(x_1', x_1, x_2)$ (in descending order)
4: If k distributions cannot be identified, assume they are from a single distribution set $R_i(x_1', x_1, x_2)$ as the mean of the distribution and $p_i(x_1', x_1, x_2) = p_i(x_1', \bar{x}_1, \bar{x}_2)$ where $x_1, x_2 \neq \bar{x}_1, \bar{x}_2$. In case this assignment fails, set $p_i = 1/k$ for all k.
5: Define the action space for each player: $\mathcal{F}_1 = \{f : X_1' \to X_1\}, \quad \mathcal{F}_2 = \{g : [k] \to X_2\}$
6: Construct a payoff matrix where each cell corresponds to a pair of functions $(f, g) \in \mathcal{F}_1 \times \mathcal{F}_2$. For each pair (f, g), compute the payoff $\sum_{X_1', i} P(X_1') p_i(x_1', f(x_1'), g(i)) R_i(x_1', f(x_1'), g(i))$
7: $(f^*, g^*) \leftarrow$ `Find-CNE` on constructed payoff matrix without the action spaces $\mathcal{A}_2^1, \mathcal{A}_2^1 \cup \mathcal{A}_2^2$
8: **Return:** Strategy f^*.

multiple layer selection games and CNEs. Next, we look at how causal strategies compare with other strategies. NE$(\Gamma(A); L_\Gamma)$ is the NE payoff with action space A as chosen in L_Γ.

Theorem 3.6 (Dominance of causal strategies)**.** *Let Γ be a CNFG with CNE payoff μ^* and L_Γ be its layer selection game with NE strategy s^*. If s^* is a pure strategy NE and $A_i^* = supp(s_i^*)$, $\mu^* \geq \texttt{NE}(\Gamma(A_i, A_{-i}^*); L_\Gamma)$ for all $A_i \in \{A_i^1, A_i^2, A_i^1 \cup A_i^2, A_i^3\}$ and $i \in [n]$.* □

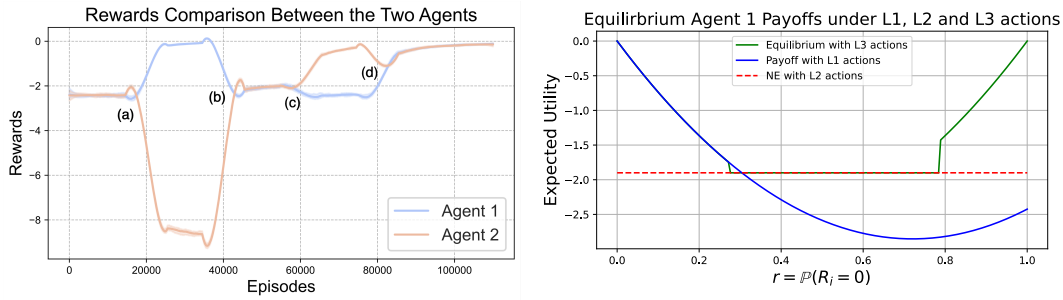
In other words, Thm. 3.6 guarantees that if the layer selection game L_Γ admits a pure strategy NE, no agent benefits by unilaterally switching to a different PCH reasoning layer. Consider Fig. 4a, which shows the layer selection game for the game in Fig. 2b – if Player 1 follows L_2 policies and Player 2 follows L_1 and L_2 policies, neither has an incentive to switch to a different layer of PCH. This leads to an interesting insight – CNE payoff of Γ is thus $(0, 0)$, while the NE payoff of Γ with L_3 actions is $(-1.5, -1.5)$ and that with interventions is $(0, 0)$. In contrast, Fig. 4b, corresponding to the CPD in Ex. 3.1, has a pure strategy NE at $(\mathcal{A}^3, \mathcal{A}^3)$, indicating both players should adopt L_3 policies. This is consistent with Fig. 3 resulting in a payoff $(0, 0)$ while NE payoff with interventions is $(-1.9, -1.9)$.

4 Learning Causal Nash Equilibrium

In this section, we introduce two algorithms for computing the CNE in CNFGs. First, we present `Find-CNE` (Alg. 1), which applies when the payoff matrix is common knowledge, as in SGT. Then, we propose `Ctf-Nash-Learning`, which learns the payoff matrix under partial observability.

We begin with the setting where the action spaces and corresponding payoffs of the CNFG Γ are known to both agents (as in SGT). For example, if Player 1 has access to L_3 and Player 2 to L_2, both are aware of the payoffs for all combinations of actions within those spaces. We introduce `Find-CNE` (Alg. 1), which implements the ideas presented in Sec. 3. The algorithm first constructs the layer selection game L_Γ corresponding to Γ (step 2). and then computes its NE strategy (step 3). Any action space that occurs with nonzero probability in the NE strategy is used for CNE, or else discarded. Step 4 computes the NE of the projection of Γ with the restricted action space.

However, such game dynamics may not be common knowledge. If the agents are learning the payoff matrix through exploration, they may be able to observe only the other agents' executed actions, but not their intuitions. To this end, we propose `Ctf-Nash-Learning` (Alg. 2), an algorithm that learns the payoff matrix in two-player CNFGs, where both agents have access to L_3 policy space. We assume that during exploration or learning phase, both players are playing `Ctf-RCT` [2] and collect



(a) (b)

Figure 5: (a) Change in payoffs of the players in Causal Prisoner's Dilemma move up the layers of PCH. Transitions (a), (b), (c) and (d) corresponds to the ones indicated in Fig. 3 (b) Equilibrium Player 1 Payoffs with L_1, L_2 and L_3 action spaces under two conditions.

the dataset $(x_1', x_1, x_2, \mathbf{y})$ (for player 1), where x_1' is the intuition of player 1, x_1 and x_2 are the actions executed by players 1 and 2, respectively, and $\mathbf{y}$ is the reward tuple. The agents do not know the solution to the layer selection game or the optimal layer in which to play.

For a fixed (x_1', x_1, x_2), the outcome $\mathbf{y}$ is sampled from the mixture $\sum_{x_2'} P(x_2' \mid x_1')P(\mathbf{y}_{x_1,x_2} \mid x_2', x_1')$. Step 3 recovers the means and weights of the mixture, which correspond (up to permutation) to $P(x_2' \mid x_1')$ and $E[\mathbf{Y}_{x_1,x_2} \mid x_1', x_2']$. In the CPD example, we identify $p_1(x_1', x_1, x_2) \approx 0.6$ and $p_2(x_1', x_1, x_2) \approx 0.4$ for all (x_1', x_1, x_2), matching $P(U_1 = 0)$ and $P(U_1 = 1)$. Examples of sample means include $R_1(0,0,0) = (-1.5, -1.5)$ and $R_2(0,0,0) = (-1, -1)$, corresponding to expectations conditioned on $X_2' = 0$ and $X_2' = 1$, respectively. These values can be consistently identified under certain technical assumptions (Appendix D). Step 4 addresses the degenerate cases where $\mathbf{Y}$ does not vary with intuition. Step 5 defines the agents' L_3 action spaces. In CPD, for agent i, it is $\{f(x) = x, f(x) = 0, f(x) = 1, f(x) = 1 - x\}$ corresponding to actions $\{X_i = X_i', do(X_i = 0), do(X_i = 1), X_i = 1 - X_i'\}$. However, the other agents' intuitions deduced in this manner may be a permutation of the actual intuitions X_2'. Once we have a proxy for the L_3 actions, the payoff matrix can be computed using Step 6 and the CNE strategy using `Find-CNE`. The learned probabilities, mean, and payoff matrix for CPD are shown in the Appendix D.

Theorem 4.1. *Given a two player CNFG $\Gamma = \langle \mathbb{M}, (\mathcal{A}_1^3, \mathcal{A}_2^3), \mathcal{R} \rangle$, let s^* be the NE strategy of the corresponding PCH-LSG L_Γ and $A_2 = \bigcup_{\mathcal{A} \in supp(s_2^*)} \mathcal{A}$. If $A_2 \in \{\mathcal{A}_2^2, \mathcal{A}_2^3\}$, then `Ctf-Nash-Learning` correctly learns the CNE strategy for Player 1.* □

Experimental evaluation. We empirically investigate how the behavior of the game changes when the players move across the layers of PCH. In order to simulate two agents learning, we enable them with Independent Q-Learning [41], a multi-agent RL algorithm that does not require knowledge of the other agents. The dynamics as Player 1 moves up the layers of PCH, while Player 2 remains in the previous layer is shown in Fig. 5a. This is an experimental realization of the discussions presented in Ex. 3.1 and Fig. 3. Every 20,000 timesteps, one of the agents moves up the layers of PCH, which triggers a change in payoff.

Next, we also investigate how the equilibrium payoffs change with the value of $P(R_1 = 0)$ and $P(R_2 = 0)$ for the scenario presented in Ex. 1.1. In this setting, we showed two extreme cases when $P(R_i = 0) = 0$ and $P(R_i = 0) = 1$ for $i \in \{1, 2\}$. In Fig. 5b, we show the equilibrium payoffs for different values of $P(R_i = 0) = r$ for $i \in \{1, 2\}$. Note that, for the causal prisoner's dilemma, following L_3 policy space is better than following only L_2 action space.

5 Conclusions

In this work, we examine the tension between rational and irrational decision-making through a causal lens. We introduce the Causal Prisoner's Dilemma, where rationality is optimal in one setting and being instinctive in another, despite both yielding the same game-theoretic solution. This reveals a limitation of standard methods in evaluating such choices. To address this, we propose a causal

framework that captures both rational and instinctive or biased behaviors, and show that it strictly generalizes Normal Form Games (Thm 2.11). We define counterfactual strategies and analyze equilibrium properties under these strategies (Thm 3.6). Finally, we develop algorithms to compute the Causal Nash Equilibria: `Find-CNE` (with known payoffs) and `Ctf-Nash-Learning` (learning through interaction). We hope that this framework advances the design of more robust, rational decision-making systems.

Acknowledgments

We thank Vishal Misra, Arvind Raghavan, and Mihalis Yannakakis for their thoughtful comments and discussions. This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.

[2] 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.

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

[4] 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.

[5] Kaushik Basu. The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395, 1994.

[6] Lawrence Chan, Andrew Critch, and Anca Dragan. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*, 2021.

[7] Andrew M Colman. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and brain sciences*, 26(2):139–153, 2003.

[8] J. Correa, S. Lee, and E. Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[9] Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10093–10100, 2020.

[10] Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11487–11495, 2021.

[11] Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

[12] Mauricio Gonzalez-Soto, Luis E Sucar, and Hugo J Escalante. Causal games and causal nash equilibrium. *arXiv preprint arXiv:1910.06729*, 2019.

[13] York Hagmayer and Steven A Sloman. Decision makers conceive of their choices as interventions. *Journal of experimental psychology: General*, 138(1):22, 2009.

[14] Lewis Hammond, James Fox, Tom Everitt, Alessandro Abate, and Michael Wooldridge. Equilibrium refinements for multi-agent influence diagrams: theory and practice. *arXiv preprint arXiv:2102.05008*, 2021.

[15] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. Reasoning about causality in games. *Artificial Intelligence*, 320:103919, 2023.

[16] John C Harsanyi. Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182, 1967.

[17] Nigel Howard. Paradoxes of rationality: theory of metagames and political behavior. *(No Title)*, 1971.

[18] Ronald A Howard, RM Oliver, and JQ Smith. From influence to relevance to knowledge influences diagrams, belief nets and decision analysis. *Eds. RM Oliver and JQ Smith, John Wiley & Sons Ltd*, pages 3–23, 1990.

[19] Daniel Kahneman. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9):697–720, 2003. doi: 10.1037/0003-066X.58.9.697.

[20] Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

[21] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914185.

[22] Daniel Kahneman and Amos Tversky. Choices, values, and frames. *American psychologist*, 39 (4):341, 1984.

[23] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[24] Michael Kearns, Michael L. Littman, and Satinder Singh. Graphical models for game theory. UAI'01, page 253–260, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

[25] Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

[26] Steffen L Lauritzen and Dennis Nilsson. Representing and solving decision problems with limited information. *Management Science*, 47(9):1235–1251, 2001.

[27] Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in neural information processing systems*, 33:8565–8576, 2020.

[28] George Loewenstein. The role of affect in decision making. *Handbook of Affective Sciences/Oxford UniversityPress*, 2003.

[29] John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[30] William Nichols and David Danks. Decision making using learned causal structures. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[31] Judea Pearl. *Causality*. Cambridge university press, 2009.

[32] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[33] A. Raghavan and E. Bareinboim. Counterfactual realizability and decision-making. In *The 13th International Conference on Learning Representations*, 2025. forthcoming.

[34] Arvind Raghavan and Elias Bareinboim. Counterfactual realizability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uuriavczkL.

[35] Frank P Ramsey. Truth and probability. In *Readings in formal epistemology: Sourcebook*, pages 21–45. Springer, 1926.

[36] Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

[37] Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation learning via inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

[39] Steven A Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996.

[40] Steven A Sloman and York Hagmayer. The causal psycho-logic of choice. *Trends in cognitive sciences*, 10(9):407–412, 2006.

[41] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

[42] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157): 1124–1131, 1974.

[43] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

[44] Amos Tversky and Daniel Kahneman. Causal schemas in judgments under uncertainty. In *Progress in social psychology*, pages 49–72. Psychology Press, 2015.

[45] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

[46] Peter C Wason and J St BT Evans. Dual processes in reasoning? *Cognition*, 3(2):141–154, 1974.

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

[48] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.

Supplementary Material

A Preliminaries and Background **14**

A.1 Structural Causal Models and PCH . . . 14

A.2 Counterfactual Randomization . . . 15

A.3 Normal Form Games and Nash Equilibrium . . . 16

A.4 Paradox of Rationality . . . 18

A.5 L_1, L_2 and L_3 actions in single-agent systems: The greedy Casino . . . 19

A.6 Graphical Models and Game Theory . . . 20

A.7 Notes on Hammond et al. [15] . . . 21

B Proofs **24**

B.1 Proof of Theorem 2.11 . . . 24

B.2 Proof of Theorem 3.5 . . . 24

B.3 Proof of Theorem 3.6 . . . 24

B.4 Proof of Theorem 4.1 . . . 24

C Discussion of Causal Games & Information sources **25**

C.1 Standard versus Causal Normal-Form Games . . . 25

C.2 Correlated Equilibrium . . . 26

C.3 Bayesian Games . . . 28

C.4 CNFG and Correlated Equilibrium . . . 30

C.5 CNFG and Bayesian Games . . . 30

D Additional Examples and Discussion **32**

D.1 SCM for Table 2b . . . 32

D.2 Assumptions for Alg. 2 . . . 32

D.3 Ctf-Nash Learning on Causal Prisoner's Dilemma . . . 33

D.4 Forgetting in Prisoner's Dilemma . . . 34

E FAQ **34**

A Preliminaries and Background

A.1 Structural Causal Models and PCH

Structural Causal Models is a general class of data-generating models [31, 2] that allows three types of distributions based on three levels of interaction with the system: observational, interventional, and counterfactual. First, we will give the formal definitions of these concepts and the hierarchical relation among them, known as Pearl Causal Hierarchy (PCH). Our presentation mostly follows Bareinboim et al. [4].

Definition A.1 (Structural Causal Models)**.** A structural causal model $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where

- $\mathbf{U}$ is a set of background variables, also called exogenous variables, that are determined by factors outside the model;
- $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, that are determined by other variables in the model — that is, variables in $\mathbf{U} \cup \mathbf{V}$.
- $\mathcal{F}$ is the set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from (the respective domains of) $U_i \cup Pa_i$ to V_i, where $U_i \subset \mathbf{U}, Pa_i \subseteq \mathbf{V} \setminus V_i$, and the entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$, that is for each $i = 1, 2, \ldots, n$, we have $v_i \leftarrow f_i(pa_i, u_i)$;
- $P(\mathbf{U})$ is the distribution over $\mathbf{U}$.

One way to visualize the dependence among the variables in the SCM is through a causal diagram, formal construction of which is given below (Def. 13, [4]).

Definition A.2 (Causal Diagram (Semi-Markovian Models))**.** Given an SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, a causal diagram G of $\mathcal{M}$ is constructed as follows:

1. add a vertex for every endogenous variable in the set $\mathbf{V}$
2. add an edge $(V_i \rightarrow V_j)$, for every $V_i, V_j \in \mathbf{V}$ and V_i occurs as an argument in $f_j \in \mathcal{F}$.
3. add a bidirected edge $(V_i \leftarrow \ldots \rightarrow V_j)$ for every $V_i, V_j \in \mathbf{V}$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or the corresponding functions f_i, f_j share some $U \in \mathbf{U}$ as an argument.

Next, we define three types of distributions corresponding to distinct modes of interaction with an SCM: the L_1 (observational), L_2 (interventional), and L_3 (counterfactual) distributions (Defs. 2, 5, and 7 in [4]).

Definition A.3 (L_1 valuation)**.** An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ defines a joint probability distribution $P^{\mathcal{M}}(\mathbf{V})$ such that for each $\mathbf{Y} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}) = \sum_{\mathbf{u} | \mathbf{Y}(\mathbf{u}) = \mathbf{y}} P(\mathbf{u}) \tag{6}$$

Before we define L_2 evaluations, we need to understand interventional SCMs. Let $\mathcal{M}$ be an SCM and $\mathbf{x}$ be an assignment to $\mathbf{X} \subseteq \mathbf{V}$. Then the interventional SCM $\mathcal{M}_{\mathbf{x}}$ is the 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle$, where $\mathcal{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} \leftarrow \mathbf{x}\}$. This operation is also known as the $do(\mathbf{x})$ operation.

Definition A.4 (L_2 valuation)**.** An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family a joint probability distributions over $\mathbf{V}$, one for each intervention $\mathbf{x}$. For each $\mathbf{Y} \subseteq \mathbf{X}$,

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\mathbf{u} | \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}} P(\mathbf{u}) \tag{7}$$

where $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$

Such an operation, where the values of random variables $\mathbf{X}$ are set to constant values $\mathbf{x}$, is known as hard interventions. Conditional or stochastic interventions can be defined similarly [9]. Let $\sigma_{\mathbf{X}} = \{\sigma_X\}_{X \in \mathbf{X}}$ be the set of soft-interventions on the variables $X \in \mathbf{X}$. Given $\sigma_{\mathbf{X}}$, the new

Algorithm 3 `Ctf-RCT`: Counterfactual Randomized Controlled Trials in MAB

Input: domain of actions $\mathcal{D}(X)$, total number of trials $N \in \mathbb{N}$
for $t = 1, 2, \ldots$ **do**
  Perceive intended action $X^{(t)}$ and store it.
  if $t \leq N$ **then**
    Sample realized action
$$X'^{(t)} \sim \text{Unif}\big(\mathcal{D}(X)\big).$$
  else
    Set
$$X'^{(t)} = \arg\max_{x} \widehat{\mathbb{E}}^{(N)}\big[Y_{X \leftarrow x} \mid X = X^{(t)}\big].$$
  end if
  Perform $\text{do}\big(X'^{(t)}\big)$ and receive reward $Y^{(t)}$.
end for

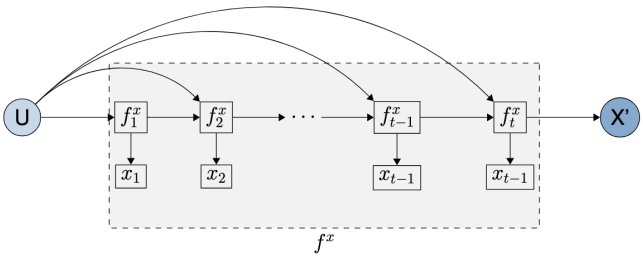


Figure 6: Illustration of decision flow f_X

model $M_{\sigma_{\mathbf{X}}}$ is defined as $\langle \mathbf{V}, \mathbf{U} \cup \mathbf{U}'_{\mathbf{X}}, \mathcal{F}', P(\mathbf{U} \cup \mathbf{U}'_{\mathbf{X}}) \rangle$, where $\mathbf{U}'_{\mathbf{X}} = \{\mathbf{U}'_X\}_{X \in \mathbf{X}}$ and $\mathcal{F}' = \{\mathcal{F} \setminus \{f_X\}_{X \in \mathbf{X}}\} \cup \{f'_X\}_{X \in \mathbf{X}}$. The distribution $P(\mathbf{V}; \sigma_{\mathbf{X}})$ can then be computed as $P(\mathbf{V})$ in $M_{\sigma_{\mathbf{X}}}$.

Next, we move on to the L_3 distributions, where we ask questions of the form "Given that the patient died without the treatment, would they be alive if they were given the treatment?". The first thing to note here, is that this query is not the same as treatment effect, that is $E[Y_{x=1}] - E[Y_{x=0}]$, where we are taking the difference between the average effects of giving the treatment and not giving the treatment. On the other hand, in the counterfactual question, we are asking the question if it would have helped for the same individual. Now, the difficulty of this problem, lies in the fact, that the patient was already denied treatment and died, and it is not practical to go back in time and give them the treatment. Mathematically, if Y is the variable that denotes whether is the patient is alive and X be the variable that the patient was given the treatment, we can write the above question as $P(Y_{x=1} = 1 \mid X = 0, Y = 0)$. Now, we provide a formal definition on how to compute counterfactual queries, given an SCM.

Definition A.5 (L_3 valuation)**.** An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $\mathbf{y_x}, \ldots \mathbf{z_w}$ for $\mathbf{Y}, \mathbf{Z}, \ldots, \mathbf{W}, \mathbf{X} \in \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y_x}, \ldots \mathbf{z_w}) = \sum_{\mathbf{u} \mid \mathbf{Y_x}(\mathbf{u}) = \mathbf{y}, \ldots \mathbf{Z_w}(\mathbf{u}) = \mathbf{z}} P(\mathbf{u}) \tag{8}$$

The collection of observational (L_1), interventional (L_2) and counterfactual (L_3) are together called the PCH.

A.2 Counterfactual Randomization

In practice, interacting through L_3 or counterfactual layer of the Pearl Causal Hierarchy can be extremely nontrivial. To this end, [3] introduces a novel form of randomization to interact through the Layer 3 of PCH. The challenge stems from the observation that agents may consider various alternatives during the deliberation process and change their opinion about the best course of action. For example, if the agent initially considers $X = x$, and then reconsiders and changes to $X = x'$, is

this counterfactual action where the natural intuition was x and the performed action would be x'? What if the agent reconsiders their decision again and changes it to x; is the agent acting against their intuition? What is the intuition x or x'. Counterfactual randomization [3, 2] addresses this concern.

The main idea is that the agent may consider many options during the deliberation process, but only the final choice matters. Consider the deliberation process shown in Fig. 6; at time step $T = 1$, the agent intends to play $X = x_1$ but reconsiders, thinking it might be sub-optimal, and decides to switch to $X = x_2$ instead, where $x_1 \neq x_2$. As time passes, the agent may realize that $X = x_{t-1}$ was not ideal and switch to an alternative, $X = t$. Ultimately, the final decision defines the intuition of the agent, regardless of the path taken to reach it. In practice, the agent could also in this reasoning process forever without ever reaching a decision.

This challenge calls for novel counterfactual machinery to allow for the counterfactual interaction following layer 3. Bareinboim et al. [3] introduces counterfactual randomization in which an agent is interrupted just before the execution of the choice, the choice being taken as the natural intuition and the final action executed based on this intuition and flip of a coin. Further, Bareinboim et al. [2] also introduces `ctf-RCT`, where an intuition is observed and then an action is chosen at random for execution. This allows us to sample from counterfactual distributions of the form $P(Y_x \mid x')$ $E[Y_x \mid x']$, where Y is the outcome, x is the intervened value and x' is the intuition, measured just before the decision. The algorithm is shown in Alg. 3. For more details on this procedure in the single-agent setting, please refere to Bareinboim et al. [2, Sec. 7]

A.3 Normal Form Games and Nash Equilibrium

In many settings – from economics and political science to computer science and biology – multiple decision-makers interact strategically, each trying to achieve the best possible outcome for themselves. A *normal-form game* provides a compact way to model such one-shot interactions, and the concept of *Nash equilibrium* captures the idea of a stable outcome where no individual can benefit by unilaterally changing their choice. In this section of the appendix, we walk through these ideas step by step, illustrating them with the classical Prisoner's Dilemma and with the intent of contrasting this later on with other variations and approaches.

A.3.1 What is a Normal-Form Game?

Intuitively, a normal-form game asks:

> *"If each player picks an action simultaneously, how do their combined choices determine everyone's payoffs?"*

This question leads to the following definition of a game:

Definition A.6 (Normal-Form Game)**.** A finite n-player normal-form game is a tuple:

$$G \;=\; \langle N,\; A,\; u \rangle$$

where

- $N = \{1, 2, \ldots, n\}$ is the set of players.
- $A = A_1 \times A_2 \times \cdots \times A_n$, with each A_i a finite set of actions available to player i. An element $a = (a_1, a_2, \ldots, a_n)$ is called an *action profile*.
- $u = (u_1, u_2, \ldots, u_n)$ is a collection of payoff functions, one per player:

  $$u_i : A \;\longrightarrow\; \mathbb{R}, \quad a \;\mapsto\; u_i(a).$$

  Given a profile a, $u_i(a)$ tells us how much player i "earns" (or how happy they are) under that combination of actions.

To summarize, in Normal Form Games:

- Each player i simultaneously chooses an action $a_i \in A_i$.
- Once all choices $a = (a_1, \ldots, a_n)$ are made, each player i receives payoff $u_i(a)$.

Despite their simplicity, Normal-Form Games are extremely powerful in their expressive power and many other richer representations, such as Extensive Form and Bayesian Games, can be reduced to a Normal Form equivalent. Next, we look at how agents can and should behave in a normal form games.

A.3.2 Mixed Strategies and Best Responses

Rather than committing to a single action, players may *randomize* over their options. A *mixed strategy* for player i is simply a probability distribution over A_i. Denote by

$$S_i = \Delta(A_i)$$

the set of all such distributions, and by $S = S_1 \times \cdots \times S_n$ the collection of all players' mixed strategies.

Given that the other players use some mixed-strategy profile $s_{-i} \in S_{-i}$, player i will choose a distribution $s_i \in S_i$ to maximize their expected payoff

$$u_i(s_i,\, s_{-i}) \;=\; \sum_{a \in A} \big[s_i(a_i) \times s_{-i}(a_{-i})\big]\; u_i(a).$$

Definition A.7 (Best Response)**.** A mixed strategy $s_i^* \in S_i$ is a *best response* to opponents' strategy profile s_{-i} if

$$u_i\big(s_i^*,\, s_{-i}\big) \;\geq\; u_i\big(s_i,\, s_{-i}\big) \quad \text{for every } s_i \in S_i.$$

In other words, s_i^* gives player i the highest possible expected payoff, assuming the others stick to strategy s_{-i}.

The notion of best response will play a key role in understanding agent's behavior.

A.3.3 Nash Equilibrium

A Nash equilibrium is a collection of strategies – one per player – such that each player's choice is a best response to everyone else's. No one can gain by deviating alone. Such a notion gives a concept of stability in a game, or a type of a solution.

Definition A.8 (Nash Equilibrium)**.** A mixed-strategy profile $s^* = (s_1^*, \ldots, s_n^*)$ is a *Nash equilibrium* if, for every player i,

$$u_i\big(s_i^*,\, s_{-i}^*\big) \;\geq\; u_i\big(s_i,\, s_{-i}^*\big) \quad \text{for all } s_i \in S_i.$$

A.3.4 Example: The Prisoner's Dilemma

To see these definitions in action, consider the *Prisoner's Dilemma*, a two-player game where each must choose either to cooperate (C) or defect (D). The actions of the player i can be written as:

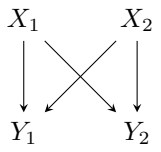
$$A_i = \{C, D\}.$$

Their payoffs are given by the following matrix (first entry is player 1's payoff, second is player 2's):

P1 \ P2	**C**	**D**
C	$-1,\ -1$	$-7,\ -0.5$
D	$-0.5,\ -7$	$-1.9,\ -1.9$

Let's walk through each prisoner's incentives:

- If player 2 cooperates (C), then player 1's payoff is
$$u_1(C, C) = -1 \quad \text{vs.} \quad u_1(D, C) = -0.5.$$
player 1 is better off defecting (D), since $u_1(D, C) > u_1(C, C)$.
- If player 2 defects (D), then player 1's payoff is
$$u_1(C, D) = -7 \quad \text{vs.} \quad u_1(D, D) = -1.9.$$
Again, player 1 prefers to defect (D), since $u_1(D, D) > u_1(C, D)$.



(a) Causal graph for Markovian Prisoner's Dilemma

Convict 2 / Convict 1	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$-1, -1$	$-7, -0.5$
$X_1 = 1$	$-0.5, -7$	$-1.9, -1.9$

(b) Payoff matrix for Prisoner's Dilemma

Figure 7: Representation of the Markovian Prisoner's Dilemma: (a) causal graph and (b) corresponding payoff matrix.

- Hence, it is better for player 1 to defect D, irrespective of what player 2 does. By symmetry, player 2 likewise always prefers D whatever player 1 does.

Thus, each player's unique *best response* to the other is to defect. When both play their best responses, we reach the profile, (D, D) which is the game's *Nash equilibrium*. Ironically, although (C, C) would yield $(-1, -1)$ total payoff (mutual cooperation), rational self-interest drives both to (D, D), giving only $-1.9 - 1.9 = 3.8$.

A.4 Paradox of Rationality

It has been observed throughout economics and behavioral game theory literature that irrationality can result in better outcomes than rational choices. One such example is the prisoner's dilemma. Both cooperating would be an irrational choice, but it results in a better payoff compared to fully rational players both of whom would choose to confess. Such irrational co-operations have also been observed in practice [7]. There has been several attempts in order to explain such irrationalities observed in human decision-making either through different models of bounded rationality, such as payoff transformations [43, 22, 23] or through alternate forms of reasoning [7].

Consider the example of the Travelers' Dilemma [5], where 2 travelers are asked to write the price of their lost item between $2-100. One with the lower value receives the lower value + $2 and one with the higher value receives lower value - $2. If an agent just tries to maximize their own reward and do not reason over others, both of them will write $100 and receive that. Now, if they do one step of reasoning, they will think "If I write $99 and my opponent writes $100 then, I will get $101 and my opponent $97". Hence, both will write $99 and get $99. The amount will decrease with more levels of reasoning. Irrational players again get higher payoffs than rational agents.

Basu [5] states that different thought processes lay behind different types of choices that people made playing a version of Traveler's Dilemma with the options ranging from 180 to 300 (pie chart): a spontaneous emotional response (choosing 300), a strategically reasoned choice (295–299) or a random one (181–294). Players making the formal rational choice (180) might have deduced it or known about it in advance. As expected, people making "spontaneous" or "random" selections took the least time to choose (as seen in experiments).

A.4.1 Causal Game Theory and Paradox of Rationality

Our proposed framework can both model and explain this gap between theory and practice. First, we consider the modeling part through the example of Prisoner's Dilemma, which we will also call the Markovian Prisoners Dilemma. The causal graph for the Markovian PD is shown in Fig. 7a. Let X_1 and X_2 be the action variables and their values 0 and 1 correspond to cooperating (C) and defect (D) respectively, and let their natural probability of cooperating be $P(X_i = 0) = 0.9$ for $i \in \{1, 2\}$. If someone is simply collecting observational data, it may happen that the agents are simply playing L_1 and hence the corresponding payoff is higher. Thus our modeling of games can model the irrational tendencies of the agents involved.

Next, comes the explaining part. Consider the scenario M_1 in Ex. 1.1, and the optimal equilibrium action, which is both the players playing L_1. Note that, this is infact the best choice from the perspective of a player, who knows how to act in L_1, L_2 or L_3. However, from an external observers point of view, these players are playing suboptimally, that is, playing C with probability 0.6 and D otherwise. From an external's observers' point of view, if the agents performed RCT, they would have a payoff as shown in Fig. 1b (bottom table), according to which playing D is the optimal strategy

with a payoff $(-1.9, -1.9)$. However, the agents, playing seemingly irrationally somehow get a payoff of $(0, 0)$, creating a paradox in the mind of the external experiment designer.

A.5 L_1, L_2 and L_3 actions in single-agent systems: The greedy Casino

The following examples illustrates the limitations of traditional decision-making and how an agent can interact with the system through the three layers of PCH as first introduced by Bareinboim et al. [3]. Consider a casino introducing two new slot machines, denoted 0 and 1. Gamblers choose machines according to two unobserved binary factors: their level of inebriation $(D \in \{0, 1\})$ and whether a machine is blinking $(B \in \{0, 1\})$. Although these factors are hidden from the agent, they influence natural behavior through the rule $X = D \oplus B$, determining the arm $X \in \{0, 1\}$ a gambler is predisposed to choose.

The casino exploits this behavioral pattern by designing reactive slot machines that adjust payouts based on these hidden variables. While ensuring that payout rates meet a government-mandated minimum of 30% when players are assigned arms randomly (e.g., RCT during inspection), the machines covertly reduce payout rates for players who follow their natural inclinations. The effective payouts are given in Table 1.

	$D = 0$		$D = 1$	
	$B = 0$	$B = 1$	$B = 0$	$B = 1$
$X = 0$	**0.10**	0.50	0.40	**0.20**
$X = 1$	0.50	**0.10**	**0.20**	0.40

Table 1: Reactive slot machine payouts: bolded entries indicate natural arm choices under the rule $X = D \oplus B$.

Note, that while players are following their natural choice, the payoff is given by

$$E[Y] = \sum_{b,d} y \cdot P(y \mid X = d \oplus b, b, d) = 0.1 \tag{9}$$

On the other hand, if the inspectors do an RCT, the payoff is given by

$$E[Y \mid do(X = x)] = \sum_{b,d} y \cdot P(y \mid X = x, b, d) = 0.3 \tag{10}$$

for any $x \in \{0, 1\}$.

However, if the agents are following opposite of their natural intuition, then the payoff will be given by

$$E[Y_{X=0} \mid X = 1]P(X = 1) + E[Y_{X=1} \mid X = 0]P(X = 0) = 0.45 \tag{11}$$

which is significantly higher than the other strategies. In fact, we can now make the following observation:

$$E[Y_x \mid x'] > E[Y_{x'} \mid x'] = E[Y \mid x'] \tag{12}$$

for any $x \neq x'$. The first term corresponds to the scenario when the gambler wants to play x', but then just before execution they choose x. The payoff of such a strategy is higher than the payoff of just playing their intuition. The term where they simply choose a single machine and play (as in RCT), $E[Y'_x]$ lies between these two strategies. In general L_3 strategies will always outperform L_1 and L_2 strategies, since both can be expressed as L_3 strategies.

In fact for single agent systems, it is always better for the agents to follow L_3 policy space, as L_3 space subsumes L_1 and L_2 policies.

$$\max_{\pi} \sum_{x} E[Y_{X=\pi(x)} \mid x]P(x) \geq \max\{\max_{a} E[Y_{X=a}], E[Y]\} \tag{13}$$

For more details, please refer Bareinboim et al. [3, 2].

A.6 Graphical Models and Game Theory

Several works have studied game theory from a graphical models perspective. The main emphasis has been on the computational advantages related to learning equilibria through probabilistic reasoning and corresponding optimization tools [25, 24]. This is a part of the growing and important literature known as algorithmic game theory Roughgarden [36]. Our approach addresses key gaps in existing models, particularly concerning the assumption of Markovianity, issues of irrationality, and multi-agent interactions.

Specifically, Kearns et al. [24] introduced *graphical games* to leverage graph structures for modeling interactions among players, making equilibrium computation more efficient when compared to standard Normal Form Games. Furthermore, Koller and Milch [25] extended influence diagrams [18, 26] to multi-agent settings, where decision nodes represent strategies, and probabilistic dependencies simplify equilibrium computations. Their framework was called Multi-Agent Influence Diagrams (MAIDs). The main goal of these works was connecting graphical models and game theory, and where somewhat silent with respect to how this relate to causality, including interventions and counterfactual reasoning.

The Structural Causal Influence Model by Everitt et al. [10] connects causality with the influence diagrams literature [18, 26]. They study certain notions found in this traditional literature, including value of information, value of control, among others. Their setting focuses on single-agent settings, whereas this paper considers multi-agent interactions, including more equilibrium analysis in scenarios where agents compete in a strategic manner. They also did not consider unobserved confounding, which is one of the key challenges in typical causal settings. Another form of causal games is proposed by Gonzalez-Soto et al. [12], which again focuses on actions as interventions and ignores the other layers of operations by a player.

Hammond et al. [14] extends Koller & Milch's MAIDs by introducing the concept of MAID subgames and proposing equilibrium refinements such as subgame perfect and trembling hand perfect equilibria. The authors establish equivalence results between MAIDs and Extensive Form Games (EFGs), highlighting the computational advantages of MAIDs in representing and solving certain classes of games. Still, despite its power, this work does not explore causal implications or counterfactual strategies, which are central to our framework. Our model explicitly integrates these aspects for deeper insights into strategic decision-making and the meaning of rationality.

Unlike the Structural Causal Games framework in Hammond et al. [15], which assumes Markovian dynamics, our model handles non-Markovian influences, including unobserved confounding that impact both actions and payoffs. We note that the assumptions required to ascertain Markovianity are inapplicable in our setting, since one of our main goals is to account for irrational behavior – where the agent acts without knowing why. In a Markovian setting, the agent knows the reasons for acting in a particular way. In fact, we model irrationality through the notion of counterfactuals and extend equilibrium concepts beyond purely rational agents, as prescribed by Nash's framework. A detailed comparison with this work is provided in the next section.

The approach proposed by Chan et al. [6] embeds irrationality in the Bellman equation under a Markovian assumption in a novel way. Our model, however, allows for general irrationality without specifying any functional constraints, which is necessary in a non-Markovian setting. The assumptions required to ascertain Markovianity are inapplicable in our setting, since one of our main goals is to account for irrational behavior – where the agent acts without knowing their reasons. Furthermore, while their focus is on a single-agent environment, ours is on multi-agent, strategic settings.

By bridging these gaps, our model provides a unified view of rational and irrational behaviors through a causal lens and rooted in first principles. It also extends graphical game-theoretic models to multi-agent systems, contributing to a more comprehensive understanding of equilibrium dynamics and rationality. Notably, while our work falls within the realm of causality, it is not primarily focused on its graphical aspects, as evident throughout the main body of the paper. As mentioned earlier, the central issue addressed here concerns the most fundamental decision-making setting and how counterfactual reasoning (and counterfactual randomization Bareinboim et al. [3]) can be leveraged to model and reconcile both irrational and rational behaviors, ultimately resolving the rationality paradox. We believe that the foundational understanding developed in this pervasive setting can be generalized to more complex games, where a graphical model and a more fine-grained structure could play a role, including for computational purposes.

A.7 Notes on Hammond et al. [15]

We note that Hammond et al. [15] also claim to unify causal modeling and game theory through the formalism of causal games and structural causal games (SCGs). In this section, we provide a critical examination of this claim and demonstrate that their approach fails to capture several essential aspects of causal modeling that are fundamental to understanding how agents reason and act in complex systems. Specifically, we identify four key limitations in their formalism: (1) the absence of the causal hierarchy and associated distributions, (2) the neglect of unobserved confounding, (3) the inability to represent L_1 actions even when extended with default functions, and (4) a flawed approach to counterfactual evaluation that misinterprets key semantic and identification issues. Each of these points is discussed in detail below.

1. Absence of the Causal Hierarchy and their distributions: One of the most basic features of causal modeling is the presence of different probability distributions induced by the collection of causal mechanisms, which is organized as three qualitatively different probability distributions – observational, interventional, and counterfactual [32, 4], and which are also known as the Pearl Causal Hierarchy (PCH). These three levels of distributions separate causal models from previously used graphical models such as Bayesian networks, and are considered a novel landmark in evaluation, estimation, representation, and decision-making in complex environments. As a consequence, an agent can interact with the system in three different ways corresponding to the three layers of PCH policies: L_1, L_2, and L_3. However, Hammond et al. [15] collapse this fundamental hierarchy into a single layer by treating all agent actions as (hard and soft) interventions, disregarding the observational (L_1) and counterfactual (L_3) levels of reasoning completely. Recall the key definition of games introduced in this work [15, Def. 22]:

Definition A.9 (SCG)**.** A (Markovian) SCG $M = (G, \theta)$ is a causal game over the exogenous and endogenous variables $\mathbf{E} \cup \mathbf{V}$ such that any deterministic parameterization of the decision variables of CPD π, the induced model with join distribution $P^{\pi}(V, E)$ is an SCM.

The authors explain that:

> "An SCG can be seen as an SCM without parameters for the decision variables. Given a policy π, we recover an SCM, as we explain in more detail below."

As a result, an SCG itself does not define a natural distribution (L_1), because the decision variables, say X do not have a natural mechanism f_X and are only determined by the agents or the policy. This precludes the possibility of modeling agents that interact in a L_1 or L_3 manner, significantly restricting the expressivity of SCGs in modeling how real-world agents reason. Still, in the formalism introduced in this paper, there is a natural distribution of the decision variables (Layer 1 in the PCH), hard/soft interventions (Layer 2), and counterfactual actions (Layer 3). Such fundamental features could not be captured by the models proposed in [15], which is illustrated in the following example.

Example A.10 (Markovian Prisoner's Dilemma)**.** Two thieves are suspected of a crime and are captured. Unfortunately, there is not enough evidence to convict them. They can now cooperate ($X = 0$) or defect ($X = 1$). The payoffs for the actions of the convicts are shown in Fig. 7b, where the numbers can be interpreted as the years they have to serve in prison. The Nash equilibrium of this game is when both players defect and the payoffs are $(-2, -2)$, where both players have no incentive to cooperate. The causal diagram for such a scenario is shown in Fig. 7a.

Consider the following scenarios – in the first one, let us call it M_1, the prisoners are more loyal and their spontaneous instinct is to cooperate ($X = 0$) with a probability of 0.9 (disregarding their utility), and in the second scenario, called M_2, the instinct is to cooperate ($X = 0$) with a probability of 0.5 and confess otherwise. Now, if the agents follow their natural instincts, their payoffs in the first scenario are

$$\mu^1_{L_1} = \sum_{x_1, x_2} \mathbf{Y} \cdot P(X_1 = x_1) P(X_2 = x_2) P(\mathbf{Y} \mid x_1, x_2) = (-1.46, -1.46)$$

and in the second scenario is

$$\mu^1_{L_1} = \sum_{x_1, x_2} \mathbf{Y} \cdot P(X_1 = x_1) P(X_2 = x_2) P(\mathbf{Y} \mid x_1, x_2) = (-2.5, -2.5)$$

If μ_{NE} is the NE payoff, we can see that

$$\mu^1_{L_1} > \mu_{\text{NE}} > \mu^2_{L_1} \tag{14}$$

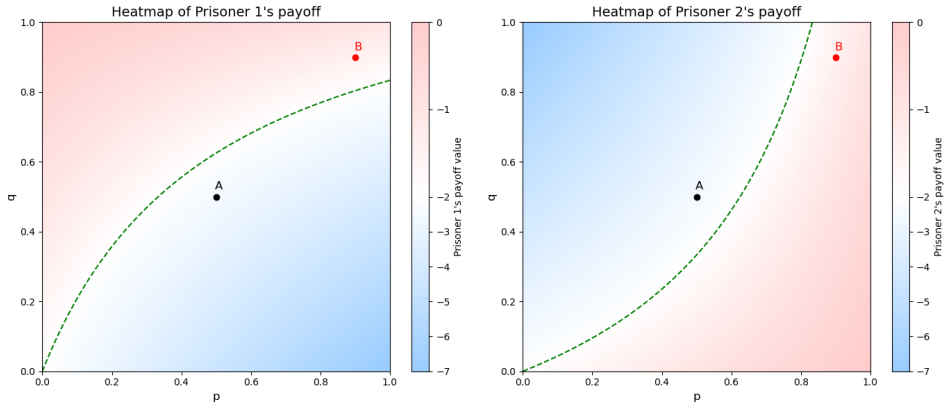


Figure 8: Payoffs of the two players as a function of their lying probability. A and B denote the payoffs when the agents follow their natural instincts in scenario M_1 and M_2, respectively.

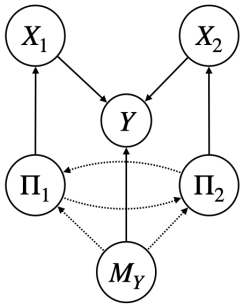


Figure 9: m-MAID is same for both M_1 and M_2 in Ex. 1.1

Now, these natural distributions cannot be represented or modeled by an SCG, where the X_1 and X_2 are determined by interventions. Hence, agents cannot act in L_1 and SCGs cannot capture the subtlety in Eq. 14. In fact, both M_1 and M_2 result in the same SCG in Fig. 7a and mechanized SCG in Fig. 9. In fact M_1 and M_2 are only two instances of infinitely many more scenarios that can happen. Fig. 8 shows how the players' L_1 payoffs change with their probability of cooperating. The scenarios M_1 and M_2 are marked as A and B in the plots. The green line denotes L_1 payoffs equal to the NE payoff. Also, since there is no concept of natural actions, L_3 policies also do not exist in SCGs. For example, in M_2 if both the prisoners decide to act against their natural instinct, then the payoffs are $\mu^2_{L_3} = (-1.46, -1.46)$ and hence $\mu^2_{L_3} > \mu_{\text{NE}}$.

2. Unobserved Confounders: One of the major challenges in real-world settings that causal inference is concerned with is the existence of unobserved confounding, variables that cannot be measured but influence both decisions and outcomes. The classic saying that "causation is not association" comes precisely because of the existence of such confounders. However, Hammond et al. [15, Sec. 2.1] does not take unobserved confounders into account.

> "In this paper, we make the simplifying assumption that all SCMs are Markovian, meaning that each variable V has exactly one exogenous parent E_V and the exogenous variables are independent."

The problem becomes even more fundamental in the context of causal game theory, as highlighted in the Causal Prisoner Dilemma (Ex. 1.1). In words, both M_1 and M_2 imply the same SCG and mechanized-MAID (Fig. 9), but entail entirely different causal analysis and decision-making strategies. For example, in M_1, following the L_1 strategies is the equilibrium strategy, and in M_2 following the L_3 strategies are the equilibrium strategies, and in both cases the equilibrium is better than the equilibrium payoffs of L_2. However, this observation is not an idiosyncrasy but rather part of a broader phenomenon, as highlighted by the following proposition.

Theorem A.11 (CNFG representation of SCG)**.** *Given an SCG corresponding to a Normal Form Game, there exist two CNFGs $\mathcal{C}_1$ and $\mathcal{C}_2$, with equilibrium payoffs μ_1 and μ_2 under L_1 and L_2 actions and Nash Equilibrium payoffs μ_{NE}, such that*

$$\mu_1 < \mu_{NE} < \mu_2 \tag{15}$$

Proof. This result follows from Thm. 2.11. Recall Def. 22 from [15], and Ex. 1.1 of this work. Hence, if we follow the proposed construction, we see there is a bijection between a normal form game and its SCG representation, which implies that given a normal form game there is a single SCG which represents the same game as the normal form game. However, Thm. 2.11 implies the existence of two CNFGs satisfying Eq. 15. Hence, CNFG represents a larger class of models than SCGs for normal form games. □

Formally, CNFGs is a strictly larger class than normal form games represented as SCG.

3. Why can't an SCG with default actions represent L_1 actions?

One may surmise that a Strategic Causal Game (SCG) could be extended to include a default action to represent an L_1 action. However, this would face two fundamental challenges:

1. **Defining the Default Action:** In CNFGs, L_1 actions are determined by nature (the SCM) through mechanisms that are entirely unknown to the agent. For instance, in Example 1.1, suppose we attempt to proxy an L_1 action in an SCG by defining a default function such as $X = U \oplus R$. But why should this specific function be chosen? It could just as well be $U \cdot R$, $U \vee R$, or any arbitrary combination of U and R. Since the agent has no knowledge of these variables or the functional form governing them, they cannot meaningfully prefer one default over another. This renders the notion of a default action indeterminate from the agent's perspective.
2. **Dependence on Unobserved Variables:** A default action that depends on unobserved variables (like U or R) inherently contradicts the assumption that these variables are unobservable to the agent. If the agent is able to use these variables as inputs in its decision-making process, then, by definition, they are no longer unobserved. This undermines the epistemic foundations on which the model is built. Put differently, this denies the possibility of unobserved confounding, a common challenge in causal inference.

These issues highlight a deeper point: SCGs and CNFGs represent fundamentally different models of agent-environment interactions. Attempting to equip SCGs with mechanisms to mimic CNFG behavior would ultimately collapse the SCG framework into that of CNFG.

4. Counterfactual Evaluation: Next, we look at the evaluation of counterfactual queries and the rationale behind computing such queries, as proposed in Hammond et al. [15]. In words, they want to answer "If we have evidence that the equilibrium π was played in the actual world, how and to what extent should that inform us of the equilibrium π' played in the counterfactual world where the values of some mechanism variables may have changed?" They claim to compute the quantity $P^{\pi'}(\mathbf{x}_I \mid z_\pi)$ through the following procedure:

- For every actual rational outcome $\pi \in \mathcal{R}(\mathcal{M} \mid z)$, update $P(\mathbf{u})$ to $P(\mathbf{u} \mid z)$ ('abduction')
- Apply the intervention I, on variables $\mathbf{Y}$, recomputing any rational responses to form π' and adding new exogenous variables as required ('action')
- Return each marginal distribution $\int_{D_\mathbf{u}} P^{\pi'}(\mathbf{x} \mid \mathbf{u}')P(\mathbf{u}')d\mathbf{u}'$ in the modified model for each counterfactual rational outcome π' ('prediction')

Note that even though this follows Pearl's algorithm in theory, it misses a fundamental practical point: neither $\mathbf{u}$ is observed, nor is $P(\mathbf{u})$ known, which makes the above evaluation impossible in most cases [31]. Formally, this procedure provides clear semantics for counterfactuals [4, Sec. 1.2], but it does not immediately imply their identification. Methods to overcome this impossibility of directly evaluating counterfactual quantities are known in the literature (see, e.g., [8, 33]). For example, one counterfactual quantity that [15] would be interested in, within our context from Example 1.1, is

$$P(Y_{X_1=1,X_2=0} = y \mid Y_{X_1=0,X_2=0} = y'), \tag{16}$$

which is known not to be identifiable from observational or interventional data without further assumptions. We note that their counterfactual evaluations do not correspond to the counterfactual actions studied in this paper. In contrast, we compute the quantity

$$\sum_x P(Y_{X_1=x',X_2=0} \mid X_1 = x)P(X_1 = x), \tag{17}$$

which SCGs cannot represent, since they do not define a natural distribution. Moreover, this quantity is counterfactually realizable and can be used in practice through counterfactual randomization [33].

B Proofs

B.1 Proof of Theorem 2.11

Consider a normal form game $\mathcal{G}$ with the action space $A = A_1 \times \ldots \times A_n$ and the utility function $u = (u_1, \ldots, u_n)$. Assume all the utilities are finite. Suppose s^* is the NE strategy and μ_{NE} is the NE payoff.

Construct an SCM as follows: $\mathbf{U} = \{U_1, \ldots, U_n\}$ and $\mathbf{X} = \{X_1, \ldots, X_n\}$, where $D_{X_i} = D_{U_i} = A_i$ and $X_i = U_i$ for all $i \in [n]$. For $i \in [n]$, $P(U_i = a_i^j) = s_i^*(a_i^j)$ where $s_i^*(a_i^j)$ is the probability of playing a_i^j by agent i in the NE strategy s_i^*. For $i = 1$

$$Y_1(a, \mathbf{u}) = u_1(a) + \mathbb{1}\{U_1 = a_1\} \cdot M \cdot (|A_1| - 1) - \mathbb{1}\{U_1 \neq a_1\} \cdot M \tag{18}$$

For all other agents $Y_i(a, \mathbf{u}) = u_i(a)$. Note that payoff for interventions is still the same as normal form games. However, if agent 1 plays L_1, the payoff can be significantly different, while for other players, the payoff remains the same as NE.

For Γ_1, suppose M is a significantly large positive number and for Γ_2, let M be a significantly large negative number, then we have for agent 1, the L_1 payoff is higher in Γ_1 and lower in Γ_2 than the NE payoff; for all other agents the payoff is the same. Hence, it follows, that

$$\mu_2 < \mu_{\text{NE}} < \mu_1$$

□

B.2 Proof of Theorem 3.5

Consider the PCH-LSG L_Γ corresponding to the CNFG Γ. Now since, L_Γ is a NFG, a mixed strategy NE exists. Let this strategy be s^*. Consider the new action space $A^* = A_1^* \times \ldots \times A_n^*$, where $A_i^* = \text{supp}(s_i^*)$. This is a fixed policy space. The PCH projection of Γ with A^* is a subgame of Γ where the action space are restricted to A^*. Now, this can be represented in Normal Form where the action space is A^*. Now, NE exists for this space. Hence, CNE exists for all CNFGs. Note that, like NFGs can have multiple NEs, CNFGs can have multiple CNEs.

B.3 Proof of Theorem 3.6

First note that $\mu^* = NE(\Gamma(A^*))$. Suppose an agent is able to change the action space from A_i^* to A_i' and improve their payoff. However, if that was true, then $NE(\Gamma(A_i', A_{-i}^*)) > NE(\Gamma(A_i^*, A_{-i}^*))$, which implies in the PCH-LSG L_Γ, agent i would be able to improve the payoff moving from A_i^* to A_i'. However, by our assumption A^* is the pure strategy NE of L_Γ, hence no such deviations are incentivised - a contradiction. Hence $\mu^* \geq NE(\Gamma(A_i', A_{-i}^*)$ for all

B.4 Proof of Theorem 4.1

First, we will show that the payoff matrix learned is a permutation of the true payoff matrix, and then find out why L_2 or L_3 payoffs will be properly learned. First note that, since the mixture is identifiable, we recover the $D_{X_2} = k$ distributions, where each of them corresponds to a value of X_2'. However the deduced values $\{\hat{x}_2^1, \ldots, \hat{x}_2^k\}$ are arranged in decreasing order of distribution, that is

$$p(\hat{x}_2^1) > \ldots > p(\hat{x}_2^n)$$

In reality, the original values of X_2' may not be so well arranged and hence $\{\hat{x}_2^1, \ldots, \hat{x}_2^k\} = h(\{x_2^1, \ldots, x_2^k\})$ where h is a permutation function.

Now, L_3 action space consists of all the functions from natural intuition X_2' to X_2. Hence the values of X_2' are essentially irrelevant and we can learn the whole table upto a permutation of the action of the second player. Since NE of Player 1 and the NE payoff remains same even with the permutation of the action space, we have that $NE(\Gamma(A_1^3, A_1^3))$ will be properly learned.

Now, L_2 action spaces are constant functions and remain invariant to permutations of X_2'. Hence, in a similar manner $NE(\Gamma(\mathcal{A}_1^2, \mathcal{A}_2^2))$ will be correctly learned, as will $NE(\Gamma(\mathcal{A}_1^3, \mathcal{A}_2^2))$ and $NE(\Gamma(\mathcal{A}_1^2, \mathcal{A}_2^3))$, and so on. By our assumption, the NE strategy of the PCH-LSG for the other agent spans over $\mathcal{A}_2^2$ and $\mathcal{A}_2^3$. Hence, the NE strategy of PCH-LSG lies on the space $\{\mathcal{A}_1^1, \mathcal{A}_1^2, \mathcal{A}_1^1 \cup \mathcal{A}_1^2, \mathcal{A}_1^3\} \times \{\mathcal{A}_2^2, \mathcal{A}_2^3\}$. Since, we are able to learn NE corresponding to each of these policies, we can correctly identify the CNE strategy.

C Discussion of Causal Games & Information sources

While the Causal Prisoner's Dilemma highlights the difficulty of cooperation – and shows that counterfactual reasoning can improve upon standard Nash-like outcomes – other strands of the literature explore strategic interactions from orthogonal perspectives, including frameworks such as Correlated Equilibrium and Bayesian Games.

The concept of Correlated Equilibrium (CE), introduced by Aumann [1], generalizes Nash Equilibrium by allowing players to coordinate their strategies through signals from an external correlation device. Unlike in Nash equilibria, where each player optimizes independently, correlated equilibria permit coordinated play, which can yield higher social welfare. This framework is particularly effective in settings where cooperation can be facilitated by signals or mediators without direct communication. In addition, CE is often easier to compute and achieves better efficiency in certain games, especially when compared to independently derived strategies. Applications include traffic routing, bargaining, multi-agent learning, and regret minimization.

Another important extension is Bayesian Games, introduced by Harsanyi [16], which address strategic interactions under incomplete information. Here, players possess private information about their own types (e.g., preferences, available actions, or payoffs) but maintain beliefs about others' types, often represented by probability distributions. This framework allows players to form and update strategies based on their beliefs. Bayesian Games naturally model scenarios involving uncertainty about the environment or about other agents, such as auctions, signaling games, contract theory, and mechanism design. From a computational standpoint, they introduce additional complexity due to the structure of beliefs and type spaces.

Comparing these frameworks to the standard Nash equilibrium setting shows how they enrich strategic analysis by incorporating coordination (in CE) and information asymmetry (in Bayesian Games). In this section, we offer a preliminary discussion on how these concepts relate to causal reasoning, and how counterfactual thinking may further enhance strategic decision-making in complex environments.

Specifically, we argue that the notion of information, as traditionally understood in the literature, is orthogonal to the causal structure captured in Causal Normal Form Games (CNFGs). This means that the causal framework can be naturally extended to incorporate sources of information available to agents. We first will revisit the standard definitions of normal form games (Sec. C.1), correlated equilibrium Sec. C.2 and Bayesian games (Sec. C.3) using formal causal language. We then introduce CNFGs with information and compare them to these classical game-theoretic frameworks (Secs. C.4 and C.5), highlighting their expressive differences and extensions.

C.1 Standard versus Causal Normal-Form Games

To ground the discussion and establish a common denonimator, we start with the definition discussed earlier, first of a causal normal form game (Def. 2.10):

Definition C.1 (Causal Normal Form Game)**.** A tuple $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$ is a Causal Normal Form Game (CNFG, for short), where

- $\mathbb{M}$ is a CMAS $\langle M, N, \mathbf{X}, \mathbf{Y} \rangle$,

- $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_n)$ is the set of policies for the n agents, where $\mathcal{A}_i \in \{\mathcal{A}^1, \mathcal{A}^2, \mathcal{A}^1 \cup \mathcal{A}^2, \mathcal{A}^3\}$,
- $\mathcal{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_n)$ is the set of reward functions.

Next, we provide the standard definition of a normal form game.

Definition C.2. A Normal-Form Game is defined as a 3-tuple $G = (N, A, u)$ where

- N: set of players.
- $A = A_1 \times \cdots \times A_n$: action space, where A_i is the set of actions available to player i.
- $u = (u_1, \ldots, u_n)$: utility functions, where $u_i : A \to \mathbb{R}$.

Whenever the policy space is constrained for the interventional layer (L_2), CNFG reduces to an NFG.

Theorem C.3. *A CNFG with L_2 actions can be converted to an NFG and vice versa with the same action space $\mathcal{A}$.*

Proof. The proof is constructive. CNFG with L_2 actions can be converted to a normal form game, simply by having the L_2 interventions as actions and $\mathbb{E}[\mathcal{R}_i(\mathbf{Y}_i(\mathbf{x}))]$ as the utility. The other way can be done as follows: define an action variable X_i for each of the n agents, where $D_{X_i} = A_i$. Then define $Y_i(x) = u_i(x)$ and $\mathcal{R}$ as a set of identity functions. □

Hence, a definition of a Normal Form Games in the causal terms would be:

Definition C.4. A Normal-Form Game is defined as a 3-tuple $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$ where

- $\mathbb{M}$ is a CMAS $\langle M, N, \mathbf{X}, \mathbf{Y} \rangle$,
- $\mathcal{A} = (\mathcal{A}_1^2, \ldots, \mathcal{A}_n^2)$ is the set of L_2 policies for the n agents,
- $\mathcal{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_n)$ is the set of reward functions.

This gives an intuitive understanding of why and how CNFGs generalize NFGs – a claim made in Thm. 2.11. For concreteness, consider the following example.

Example C.5 (Prisoner's Dilemma)**.** Consider the classical Prisoner's Dilemma game with payoffs as shown in Fig. 7b. To represent this game as a CNFG, define X_1 and X_2 as the action variables in a CMAS, and $\mathbf{Y} = Y_1, Y_2$ as the corresponding reward signals. Assume X_1, X_2 are binary, where 0 represents cooperation (C) and 1 represents defection (D). Define Y_1 and Y_2 as deterministic functions of X_1 and X_2, based on the given payoff table. For completeness, let $X_1 = U_1$ and $X_2 = U_2$, where the prior distribution over exogenous variables is $P(U_1 = 1) = P(U_2 = 1) = 0.5$. Note that this distribution is defined for formal completeness but does not affect the L_2 payoffs.

To convert this CNFG into an NFG, we focus on L_2 actions. The L_2 policy space $\mathcal{A}^2$ maps to the action values of X_1 and X_2, i.e., either 0 (C) or 1 (D). The payoff for any joint action (x_1, x_2) is given by $\mathbb{E}[\mathbf{Y} \mid do(x_1, x_2)]$, aligning exactly with the utility function in the standard Normal Form representation. □

As discussed earlier, Thm. 2.11 noted that a CNFGs is strictly more expressive than NFGs by showing two CNFGs that agree on the interventional layer but may have different equilibirums, when other layers of PCH are considered.

C.2 Correlated Equilibrium

In this section, we investigate how CNFGs can be extended to systems with information – an important step toward modeling more realistic decision-making scenarios. We begin by examining Correlated Equilibrium through a classical example known as the Battle of the Sexes.

Example C.6 (Battle of Sexes)**.** A couple of agents want to spend time together in the evening. Agent 1 wants to go to the ballet, while Agent 2 prefers a football match. Their payoffs, based on whether they go to the ballet or football, are shown in Tab. 10. The symmetric Nash equilibrium for this game occurs when both agents go to their preferred location two-thirds of the time, yielding a

Player 1 \ Player 2	$X_2 = B$	$X_2 = F$
$X_1 = B$	$2, 1$	$0, 0$
$X_1 = F$	$0, 0$	$1, 2$

Figure 10: Payoff matrix for Battle of Sexes

P1 \ P2	L_1	$X_2 = B$	$X_2 = F$
L_1	$1.875, 1.875$	$0.875, 1.375$	$1.375, 0.875$
$X_1 = B$	$1.375, 0.875$	$2, 1$	$0, 0$
$X_1 = F$	$0.875, 1.375$	$0, 0$	$1, 2$

Figure 11: Payoff matrix for Battle of Sexes with L_1 and L_2 actions

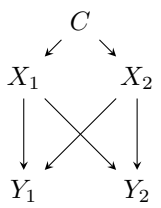


Figure 12: Battle of Sexes with Information from a coin toss

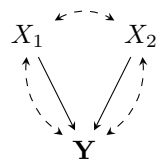


Figure 13: Causal Diagram for Battle of Sexes with unobserved confounding

joint payoff of $(0.75, 0.75)$. We can represent this Normal Form Game as a CNFG with L_2 actions. The causal diagram corresponding to the SCM is shown in Fig. 12. The actions of the agent i in this SCM is given by $do(X_i = B)$ and $do(X_i = F)$ for $i \in \{1, 2\}$.

Suppose the players now have access to a coin and observe the outcome of the toss, H, T, making decisions accordingly. Assume they follow the strategy $\{H \to B, T \to F\}$ – that is, if the coin shows heads, they go to the ballet; if it shows tails, they go to the football match. Note that this is an equilibrium strategy, as neither player has an incentive to deviate from it. The resulting equilibrium payoff is $(1.5, 1.5)$. Such an equilibrium is called a *correlated equilibrium*, and it is superior to the Nash equilibrium.

We can also represent this graphically in the causal diagram. In Fig. 12, a new variable C can be introduced, which takes two values $\{H, T\}$, each with probability 0.5. Now, since the outcome of the coin is available to the two agents, they can condition their policies on the outcome of this coin. So, the new policy space of the agents would be a mapping from outcome of coin to the show they want to attend, that is $A_i : \{H, T\} \to \{B, F\}$. If we are talking of mixed strategy, the policy of the agent is given by the distribution $\pi_i(\cdot \mid C)$ over the values $\{F, B\}$. Thus, we can simply represent additions of random variables in a correlated equilibrium as new variables in the causal model (also known as confounders).

We now formally define Correlated Equilibrium using causal language.

Definition C.7 (Correlated Equilibrium)**.** Given a CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}^2, \mathcal{R} \rangle$ with policy space restricted to L_2, a correlated equilibrium $(\mathbf{S}, P(\mathbf{S}), \pi)$ is a tuple, where $\mathbf{S} = (S_1, \ldots, S_n)$ is a tuple of random variables with distribution $P(\mathbf{S})$ and $\pi = (\pi_1, \ldots, \pi_n)$ is the set of mappings $\pi_i : D_{S_i} \to \mathcal{A}_i^2$ and for each agent i and every other mapping π_i',

$$\begin{aligned} &\sum_{\mathbf{s} \in D_{\mathbf{S}}} P(\mathbf{s}) \mathcal{R}_i(\mathbf{Y}_{i[X_1 = \pi_1(S_1), \ldots, X_i = \pi_i(S_i), \ldots X_n = \pi_n(S_n)]}) \\ &\geq \sum_{\mathbf{s} \in D_{\mathbf{S}}} P(\mathbf{s}) \mathcal{R}_i(\mathbf{Y}_{i[X_1 = \pi_1(S_1), \ldots, X_i = \pi_i'(S_i), \ldots X_n = \pi_n(S_n)]}) \end{aligned} \tag{19}$$

As observed in the previous example, it is possible to represent the correlated equilibrium as a NE of a CNFG with information. Next, we formally define CNFG where the agents have access to pieces of information (possibly shared) before acting.

Definition C.8 (CMAS with States)**.** A Causal Multi-Agent System (CMAS) with states is a tuple $\langle M, N, \mathbf{X}, \mathbf{S}, \mathbf{Y} \rangle$, where $M : \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P \rangle$ is an SCM and

- N is the set of n agents,

Police \ Suspect	$S=1$	$S=0$
$P=1$	$0,0$	$2,-2$
$P=0$	$-2,-1$	$-1,1$

Figure 14: Payoff when suspect is criminal (that is $T=1$).

Police \ Suspect	$S=1$	$S=0$
$P=1$	$-3,-1$	$-1,-2$
$P=0$	$-2,-1$	$0,0$

Figure 15: Payoff when suspect is civilian (that is $T=0$).

- $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)$ is the ordered set of action nodes with $\mathbf{X}_i,\mathbf{X}_j\subset\mathbf{V}$ for $i,j\in[n]$ and $\mathbf{X}_i\cap\mathbf{X}_j=\emptyset$ if $i\neq j$,
- $\mathbf{S}=(\mathbf{S}_1,\ldots,\mathbf{S}_n)$ is the ordered set of context nodes $\mathbf{S}_i\subset\mathbf{V}$ for the agent i for $i\in[n]$, and for all $i\in[n]$, $\mathbf{S}_i\notin De(\mathbf{X}_i)$
- $\mathbf{Y}=(\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$ is the ordered set of reward signals, with $\mathbf{Y}_i\subseteq\mathbf{V}$ for all $i\in[n]$. □

Once we have introduced the notion of a CMAS to model the environment, we can consider the information available to the agents about the states.

Definition C.9 (CNFG with Information)**.** A tuple $\Gamma=\langle\mathbb{M},\mathcal{A},\mathcal{R},\mathcal{I}\rangle$ is a Causal Normal Form Game (CNFG), where

- $\mathbb{M}$ is a CMAS with states $\langle N,M,\mathbf{X},\mathbf{S},\mathbf{Y}\rangle$,
- $\mathcal{A}=(\mathcal{A}_1,\ldots,\mathcal{A}_n)$ is the set of policies for the n agents, where $\mathcal{A}_i\in\{\mathcal{A}^1,\mathcal{A}^2,\mathcal{A}^1\cup\mathcal{A}^2,\mathcal{A}^3\}$,
- $\mathcal{R}=(\mathcal{R}_1,\ldots,\mathcal{R}_n)$ is the set of reward functions.
- $\mathcal{I}$ is the information available to the agents. □

The information $\mathcal{I}$ can take many forms and is introduced to make the definition more complete and general. For example, one form of information might be the distribution over states, $P(\mathbf{S})$, which helps illustrate the relationship between Correlated Equilibrium and equilibrium concepts in CNFGs with states. Other forms of information available to the agents could include interventional or counterfactual distributions. While this is outside the scope of the present paper, it presents a promising direction for future work.

Theorem C.10. *If $(\mathbf{S},P(\mathbf{S}),\pi)$ is a correlated equilibrium of NFG Γ, then, π is a NE of the CNFG with Information designed as follows:*

1. *$\mathbf{X}=(X_1,\ldots,X_n)$ are the actions and $\mathbf{Y}=u(x)$ for each combination are obtained from NFG Γ. $\mathcal{R}$ is identity for all agents.*
2. *Introduce variables $\mathbf{S}=(S_1,\ldots,S_n)$ with distribution $P(\mathbf{S})$.*
3. *Define the L_2 action space $\mathcal{A}_i^2$ for the agent i as soft interventions $\pi(X_i\mid S_i)$*
4. *$P(\mathbf{S})$ is available to the agents and the expected payoff for policies π is given by:*

Proof. The proof follows from the definition of correlated equilibrium. Since π is the policy in the correlated equilibrium, it is also the best response as per Eq. 19. Hence, if every agent is playing the best response given $\mathbf{S}$, we have each agent is playing NE policy (from Def. A.8) in the game with policies conditioned on $\mathbf{S}$. □

C.3 Bayesian Games

Here, we introduce an additional layer of complexity in the information structure through the concept of Bayesian Games. Before presenting the formal definition, we begin with an example.

Example C.11 (Sheriff's Dilemma)**.** A police officer faces an armed suspect, and they must simultaneously decide whether to shoot. The suspect could be either a criminal or a civilian, but the officer is unaware of the suspect's true identity. It is preferable for the suspect to shoot if they are a criminal

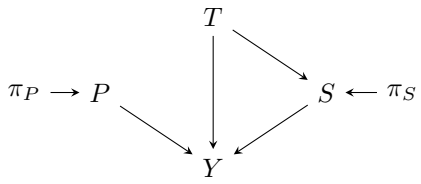


Figure 16: Causal Diagram for Sheriff's Dilemma with L_2 policy space.

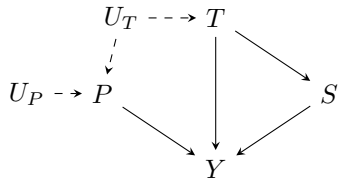


Figure 17: Causal Diagram for Sheriff's Dilemma with L_1 policy space.

and not to shoot if they are a civilian. However, in hindsight, it is better for the officer to shoot if the suspect shoots – but in reality, they must act simultaneously.

Depending on the type of the player, criminal or civilian, the payoffs corresponding to the actions of the players are shown in Fig. 14 and Fig. 15 respectively. Now, the question is how do we compute the best policy for the agents.

Let's start with the suspect's actions. If the suspect is a criminal, then shooting is a dominant strategy and if the suspect is a civilian, not shooting is the dominant strategy. However, from the perspective of the policmen, things are not so simple, since they do not know the type of the suspect. If they know that the suspect is highly likely to be criminal, then it is better for them to shoot, and otherwise not. Hence, in order to form a decision, they need to have a belief over the likelihood of someone being a criminal.

The scenarios can be represented in a corresponding causal graph shown in Fig. 16. The variable T represents the type of the suspect: $T = 0$ indicates a civilian, and $T = 1$ indicates a criminal. The variable P captures the officer's decision to shoot or not, while S denotes whether the suspect chooses to shoot. If we consider L_2 actions only, then the policy space is a soft intervention over the values of P. Finally, $Y = (Y_1, Y_2)$ represent the utilities of the officer and the suspect, respectively. The value of Y as a function of P, T and S is shown in Fig. 15 and Fig. 14. In addition, the agent should also have a belief of the likelihood of someone being criminal, that is $P(T)$.

This class of Games where there is an uncertainty about the nature of the game is called Bayesian Game. Formally, Bayesian Games can be defined as follows [16]:

Definition C.12 (Bayesian Games)**.** A Bayesian Game is a tuple $\langle N, A, \Theta, p, u\rangle$, where

- N is the set of n players indexed by i;
- $A = A_1 \times \ldots \times A_n$, where A_i is the action set available to player i;
- $\Theta = \Theta_1 \times \ldots \times \Theta_n$ where Θ_i is the type space for player i;
- $p : \Theta \to [0, 1]$ is a common prior over types;
- $u = (u_1, \ldots, u_n)$ where $u_i : A \times \Theta \to \mathbb{R}$ is the utility function for player i

Now, we formalize the idea of Bayesian Games in the causal framework.

Theorem C.13 (Bayesian Games in Causal Framework)**.** *A Bayesian Game is a CNFG with information $\Gamma = \langle M, \mathcal{A}^2, \mathcal{R}, \mathcal{I}\rangle$, where M is a CMAS with states.*

Proof. The construction follows in the same way as Normal Form Games, but now with introduction of the type variables. The CMAS contains the variables $\mathbf{X}$ corresponding to the actions A, state variables $\mathbf{S}$ corresponding to the type Θ, $\mathbf{Y}(x) = u(x)$ and $\mathcal{R}$ is the identity function. The information available to the agents is $P(\mathbf{S})$. □

Now, a best response π_i^* to a strategy profile π_{-i} in a Bayesian Game is defined as

$$BR_i(\pi_{-i}) = \arg\max_{\pi_i'} EU(\pi_i', \pi_{-i}) \tag{20}$$

Definition C.14 (Bayes-Nash Equilibrium)**.** A Bayes Nash Equilibrium is a strategy profile π such that for all i, $\pi_i \in BR(\pi_{-i})$.

		$X_2 = B$		$X_2 = F$	
		$U_2 = 0$	$U_2 = 1$	$U_2 = 0$	$U_2 = 1$
$X_1 = B$	$U_1 = 0$	$3, 0$	$3, 3$	$0, 0$	$0, 0$
	$U_1 = 1$	$0, 0$	$2, 1$	$0, 0$	$0, 0$
$X_1 = F$	$U_1 = 0$	$0, 0$	$0, 0$	$0, 3$	$0, 0$
	$U_1 = 1$	$0, 0$	$0, 0$	$3, 3$	$1, 2$

Table 2: Battle of Sexes with Unobserved Confounding

C.4 CNFG and Correlated Equilibrium

In Sec. C.2, we saw how we can represent correlated equilibrium with L_2 actions. However, SCMs can inherently represently two more layers of distribution L_1 and L_3. Using the more general action space is not only a matter of choice, but can be essential in obtaining a better payoff (that is finding the corresponding equilibrium), as illustrated in the following example.

Example C.15 (Causal Battle of Sexes)**.** Considering Ex. C.6, we note that, in reality, the decision to go to the ballet or the football game may be influenced by several external factors. For example, when a new ballet is released, agent 1 – who generally prefers football – may also want to attend the ballet. Conversely, if there is a major football event, such as the Super Bowl, agent 2 may also be happy to join agent 1 for the match. These unobserved factors can therefore influence the preferences of both players. The corresponding causal graph is shown in Fig. 13.

In addition, the importance of the event may also affect the payoffs. Let $U_1 = 0$ ($U_2 = 0$) indicate that there is a major football match (ballet performance), and $U_1 = 1$ ($U_2 = 1$) that the football match (ballet) is not particularly significant. Now, both agents' intuitions are given by:

$$X_i = \begin{cases} F & \text{if } U_1 = 0, U_2 = 1 \\ B & \text{if } U_1 = 1, U_2 = 0 \\ F \text{ or } B \text{ with equal probability} & \text{otherwise} \end{cases} \tag{21}$$

This means that if either the ballet or football performance is particularly good, the agents choose to attend that event. If both events are equally good or equally unappealing, they make the decision randomly. The payoff for such an L_1 action is $(1.875, 1.875)$. In contrast, if they instead base their decisions on signals from a coin toss, the resulting payoff is $(1.5, 1.5)$ – lower than what they would receive by following their natural intuitions.

We now define a correlated equilibrium over a general CNFG, where the agents' action spaces may correspond to L_1, L_2, or L_3 policies.

Definition C.16 (Causal Correlated Equilibrium)**.** Given a CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$, a causal correlated equilibrium $(\mathbf{S}, P(\mathbf{S}), \pi)$ is a tuple, where $\mathbf{S} = (S_1, \ldots, S_n)$ is a tuple of random variables with distribution $P(\mathbf{S})$ and $\pi = (\pi_1, \ldots, \pi_n)$ is the set of mappings $\pi_i : D_{S_i} \to \mathcal{A}_i$ and for each agent i and every other mapping π'_i,

$$\begin{aligned} &\sum_{\mathbf{s} \in D_{\mathbf{S}}} P(\mathbf{s}) \mathcal{R}_i(\mathbf{Y}_{i[X_1 = \pi_1(S_1), \ldots, X_i = \pi_i(S_i), \ldots X_n = \pi_n(S_n)]}) \\ &\geq \sum_{\mathbf{s} \in D_{\mathbf{S}}} P(\mathbf{s}) \mathcal{R}_i(\mathbf{Y}_{i[X_1 = \pi_1(S_1), \ldots, X_i = \pi'_i(S_i), \ldots X_n = \pi_n(S_n)]}) \end{aligned} \tag{22}$$

The definition is similar to that of a classical correlated equilibrium, except that the equilibrium is for a CNFG with action spaces that can span across any layer of the PCH.

C.5 CNFG and Bayesian Games

Similarly, in Bayesian Games, agents need not be restricted to layer L_2, as L_1 and L_3 policies may potentially yield higher rewards, as illustrated next.

Example C.17 (Causal Sheriffs Dilemma)**.** Consider Ex. C.11. In reality, the situation may not be so simple or well-defined. Unobserved factors might influence both the officer's assessment of the

Action Space	Payoff depends on agents' actions	Agents act based on a signal	Agents act based on type and has belief about the types
L_2	Normal Form Games (Def. C.2)	Correlated Equilibrium (Def. C.7)	Bayesian Games (Def. C.13)
L_1, L_2, L_3	Causal Normal Form Games (Def. 2.10)	Causal Correlated Equilibrium (Def. C.16)	Causal Bayesian Games (Def. C.18)

Figure 18: Comparison of different representations with information and action spaces

suspect and the suspect's decision to shoot. For instance, a suspect's background might affect both their likelihood of being a criminal and their behavior. A well-trained officer might intuitively discern such subtle cues and make a quick judgment about whether to shoot. An untrained officer, on the other hand, may lack this ability and be more prone to error. This creates unobserved confounding between the suspect's identity and the officer's tendency to shoot. In other words, the officer may not be able to articulate why they want – or do not want – to shoot, but their instinct carries information about their internal state.

Consider two scenarios, M_1 and M_2, that induce the same causal diagram shown in Fig. 17. In M_1, the officers are well-trained; in M_2, they are not. In both scenarios, let an adverse background be denoted by the variable $U_T = 1$, with $P(U_T = 1) = 0.1$. Suppose the suspect is a criminal, that is, $T = 1$ if and only if they come from an adverse background. This background may influence the suspect's behavior, which in turn can influence the officer's decision to shoot. In the causal diagram, this pathways are represented by the dashed arrows.

Further, in scenario M_1, the officer is able to pick up on these non-verbal cues, and their probability of shooting is given by $P(P = 1 \mid U_T = 1) = 0.9$ and $P(P = 0 \mid U_T = 0) = 0.9$. In the second scenario, M_2, the officer almost always makes mistakes, and their probability of shooting is given by $P(P = U_T) = 0.1$. The payoffs $\mathbf{Y} = (Y_1, Y_2)$, as a function of P, T, and S, are shown in Tables 14 and 15.

Now suppose Congress wants to recommend a new policy by passing a law that determines whether officers should shoot or not. They disregard the officers' natural intuitions entirely and compute the Bayesian Nash Equilibrium (BNE) of the game induced by the model, concluding that it is better if the officer does not shoot at all. The expected payoff for the officer under the BNE is therefore given by:

$$\mu_{\text{BE}} = -2 \cdot 0.1 = -0.2 \tag{23}$$

However, if the law is not implemented, then in M_1 and M_2, the expected L_1-payoff of the policeman are respectively

$$\mu_1 = -0.11, \quad \mu_2 = -0.99 \tag{24}$$

This implies $\mu_2 < \mu_{\text{BE}} < \mu_1$, indicating that, even though both SCMs induce the same Bayesian game, implementing the law would be harmful in M_1, while beneficial in M_2.

In essence, this is similar to the scenarios illustrated in Ex. 1.1. Now, we can rewrite the definition of Causal Bayesian Games without restricting ourselves to L_2 layer.

Definition C.18 (Causal Bayesian Games)**.** A Bayesian Game is a CNFG with information $\Gamma = \langle M, \mathcal{A}, \mathcal{R}, \mathcal{I} \rangle$, where M is a CMAS with states.

Note that even if two CMASs coincide on their L_2 distributions, they may differ in their L_1 and L_3 distributions – particularly in the corresponding L_1 and L_3 actions. This is formalized below.

Theorem C.19. *Given a Bayesian Game, there exists two Causal Bayesian Games Γ_1 and Γ_2 with expected L_1 payoffs μ_1 and μ_2 and Bayes-Nash Equilibrium (BE) payoffs μ_{BE}, such that*

$$\mu_2 \leq \mu_{BE} \leq \mu_1 \tag{25}$$

		$X_2 = 0$		$X_2 = 1$	
		$U_2 = 0$	$U_2 = 1$	$U_2 = 0$	$U_2 = 1$
$X_1 = 0$	$U_1 = 0$	$-2, 2$	$-2, -6$	$-2, -6$	$-2, 2$
	$U_1 = 1$	$2, -2$	$-4, 0$	$-4, 0$	$2, -2$
$X_1 = 1$	$U_1 = 0$	$2, -2$	$-4, 0$	$-4, 0$	$2, -2$
	$U_1 = 1$	$-2, 2$	$-2, -6$	$-2, -6$	$-2, 2$

Table 3: Y_1, Y_2 as a function of U_1, U_2, X_1, X_2 for SCM in Table 2b

Proof. The construction is similar to the one in the proof of Thm. 2.11 □

The above discussion illustrates the fact that information structure and actions based on layers of PCH are orthogonal to each other. We can have one without the other and even both of them in the same model, without compromising the other. The summary of the axis can be shown in Fig. 18.

D Additional Examples and Discussion

D.1 SCM for Table 2b

Consider the SCM with $\mathbf{U} = \{U_1, U_2\}, \mathbf{X} = \{X_1, X_2\}$ and $\mathbf{Y} = \{Y_1, Y_2\}$. The domains of U_1, U_2, X_1 and X_2 are $\{0, 1\}$. $P(U_1 = 0) = P(U_2 = 0) = 0.5$. $X_1 = U_1$ and $X_2 = U_2$. $\mathbf{Y}$ as a function of U_1, U_2, X_1, X_2 are shown in Table 3.

The action space available to Player 1 and Player 2 are $\mathcal{A}^3$ and $\mathcal{A}^1 \cup \mathcal{A}^2$ respectively.

D.2 Assumptions for Alg. 2

For the algorithm to work, we will make the following assumptions. Assume that the learning is from the perspective of Player 1.

Assumption D.1 (Identifiability of Mixture)**.** Let $\mathbf{Y}_{x_1,x_2} \mid x'_1, x'_{2i} \sim \phi_i$, for $i \in k$, where ϕ_i is a distribution dependent on x'_1, x_1, x_2 and $k = |D(X)|$. We assume that the distributions are such that their mean and weights are identifiable from their mixture upto a permutation of the i's:

$$\sum_{i=1}^{k} p_i \phi_i(x'_1, x_1, x_2) \tag{26}$$

or, the distributions are same for all $i \in [k]$.

Next, we show some example distributions and conditions that satisfy the above assumption.

Example D.2 (Deterministic Function)**.** Consider the case when $P(\mathbf{Y}_{x_1,x_2} \mid x'_1, x'_2)$ has all its mass on a single point. In addition, assume that

$$E[\mathbf{Y}_{x_1,x_2} \mid x'_1, x'_{2i}] \neq E[\mathbf{Y}_{x_1,x_2} \mid x'_1, x'_{2j}]$$

for $i \neq j$. Then, for each (x'_1, x_1, x_2) we will get k distinct values of $\mathbf{Y}$, and we can map each $(\mathbf{Y}_i, x'_1, x_1, x_2)$ to a particular i and $p_i = P(\mathbf{Y}_i \mid x'_1, x_1, x_2)$ for $i \in [k]$.

Example D.3 (Gaussian Mixtures)**.** Yakowitz and Spragins [47] showed that mixture of multi-variate Gaussians are identifiable. Hence, we can get the mixing proportions and the mean of the Gaussians from the sufficient amount of data.

The next assumption ensures that Player 1 can correctly deduce the intuition of the other player from the observations.

Assumption D.4. For all assignments x'_2, x''_2 to the natural intuition of the second player $P(x'_1, x'_2) \neq P(x'_1, x''_2)$.

Note that if $P(x'_1, x'_2)$ are sampled from a continuous distribution then the assumption is true almost surely.

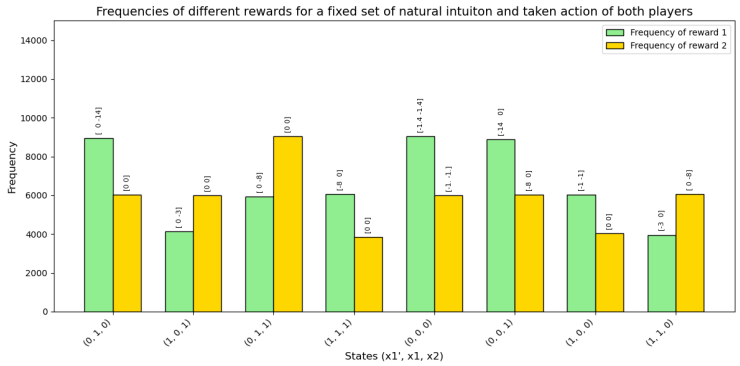


Figure 19: Frequencies of rewards observed for a particular tuple (x_1', x_1, x_2')

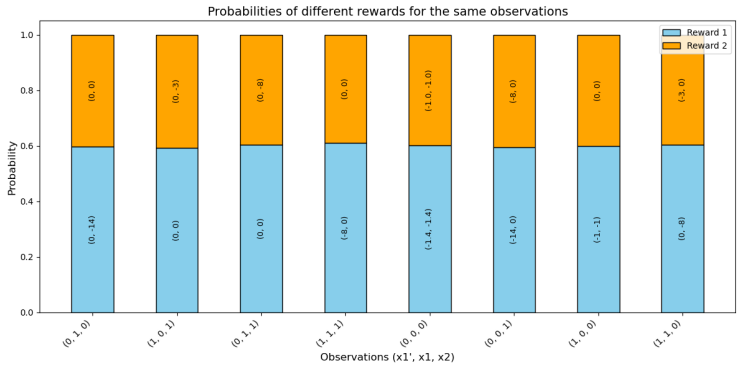


Figure 20: Probabilities of the rewards observed for a particular tuple (x_1', x_1, x_2')

Table 4: Payoff Matrix learned by Player 1 in Causal Prisoner's Dilemma

	$X_2 = X_2'$	$do(X_2 = 0)$	$do(X_2 = 1)$	$X_2 = 1 - X_2'$
$X_1 = X_1'$	$-2.443, -2.450$	$-1.218, -2.684$	$-8.892, 0.000$	$-7.668, -0.239$
$do(X_1 = 0)$	$-2.683, -1.235$	$-0.983, -0.983$	$-6.932, -0.490$	$-5.232, -0.239$
$do(X_1 = 1)$	$0.000, -8.848$	$-0.475, -6.951$	$-1.960, -1.897$	$-2.435, 0.000$
$X_1 = 1 - X_1'$	$-0.240, -7.637$	$-0.240, -5.250$	$0.000, -2.387$	$0.000, 0.000$

D.3 Ctf-Nash Learning on Causal Prisoner's Dilemma

This section shows the results of applying `Ctf-Nash-Learning` on Causal Prisoner's Dilemma. The experiment was carried out on 100K samples of $(x_1', x_1, x_2, \mathbf{y})$ when both agents were playing `Ctf-RCT`. The rewards were assumed to be deterministic, that is, $P(\mathbf{y}_{x_1,x_2} \mid x_1', x_2')$ has a point mass. Now, for each tuple (x_1', x_1, x_2) the frequencies of $\mathbf{y}$ obtained are shown in Fig. 19. For example, when (x_1', x_1, x_2') is $(0, 1, 0)$, then the reward $(0, -14)$ was observed nearly 9000 times while $(0, 0)$ was observed 6000 times, and when it is $(0, 1, 1)$, then the reward $(0, -8)$ occurs nearly 6000 times and $(0, 0)$ occurs nearly 9000 times, and so on.

From this frequency table, we can compute the probabilities as shown in Fig. 20. For instance, when (x_1', x_1, x_2') is $(0, 1, 0)$, the two values of x_2' are taken with a probability of roughly 0.6 and 0.4. The same is observed for the tuple $(0, 1, 1)$. Hence, one value of x_2' occurs with a probability of 0.6 and the other with 0.4. However, there is no way to know whether it is 0 or 1 that occurs with a probability 0.6. This results in a permuation of the actions. For example, if x_2' is correctly identified, that is $\hat{x}_2' = x_2'$ then $(\hat{x}_2' = 0, x_2 = 0), (\hat{x}_2' = 1, x_2 = 1)$ would correspond to the natural instinct or action $(X_2 = X_2')$. However, if they are not correctly identified, that is $\hat{x}_2' = 1 - x_2'$, then the same would correspond to acting against intuition $(X_2 = 1 - X_2)$.

The learned payoff matrix is shown in Table. 4. Even if the actions are permutations of the original payoff matrix, the equilibrium remains same. Hence, the algorithm will be able to find the equilibrium correctly.

The code is available at `https://anonymous.4open.science/r/CGT-NeurIPS25/`

D.4 Forgetting in Prisoner's Dilemma

As noted in Sec. 3, it is not always in the best interest of the agents to forget. Consider the classical prisoners dilemma: the choices of the action spaces are $\{C\}, \{D\}$ and $\{C, D\}$ and forget about whatever is not included in the sets. The metagame over the choice of the action spaces is shown below.

P1 \ P2	$\{C\}$	$\{D\}$	$\{C, D\}$
$\{C\}$	$-1, -1$	$-7, -0.5$	$-7, -0.5$
$\{D\}$	$-0.5, -7$	$-1.9, -1.9$	$-1.9, -1.9$
$\{C, D\}$	$-0.5, -7$	$-1.9, -1.9$	$-1.9, -1.9$

Table 5: Extended Prisoner's Dilemma with a third action CD

Note, if both the agents forget about defecting D, and plays only C, then one of the agents can move to the action space $\{C, D\}$ and get a better payoff while the other is worse of. The resultant NE of this metagame thus also turns out to be $(-1.9, -1.9)$ with action spaces $\{D\}$ or $\{C, D\}$. Thus in classical prisoners dilemma, agents do not have advantage with forgetting.

E FAQ

1. What are L_1, L_2, and L_3 actions, and why are they essential in modeling decision-making?

A: Pearl [31] and Bareinboim et al. [4] introduced a framework for studying real-world systems – ranging from experiments in medicine to analyses of climate models – using causal models such as Structural Causal Models (SCMs). An SCM induces three levels of distributions: observational (L_1), interventional (L_2), and counterfactual (L_3). One of the key results in this literature is the *Causal Hierarchy Theorem* (CHT), which states that these three levels of distributions form a strict hierarchy and do not collapse. That is, given an L_1 distribution, there may be multiple SCMs that induce the same L_1 distribution but yield different L_2 distributions. Likewise, given both L_1 and L_2 distributions, multiple models may still differ in their induced L_3 distributions [4, Thm. 1].

In decision-making systems, an agent interacts with the environment (and possibly with other agents) through its actions. If the agent does nothing and simply observes, this behavior corresponds to an L_1 action. If it disregards its intuition and performs an intervention (either hard or soft) it engages in an L_2 action. If the agent's realized action depends on what it would have done under natural circumstances (i.e., its L_1 action), then the behavior corresponds to a counterfactual, or L_3, action. These distinctions in decision-making have been studied extensively over the past decade, including in Bareinboim et al. [3, 2].

A particularly relevant work that forms the basis for our discussion in the single-agent setting is Bareinboim et al. [2], which introduces a scenario known as the Greedy Casino (reviewed in Appendix A). In this setting, there are machines that may blink and patrons who may be drunk. If a machine is blinking and a patron is intoxicated, the patron is more likely to play that machine instinctively (i.e., subconsciously). Conversely, if the patron is sober and the machine is not blinking, they tend to prefer a different option. This behavior reflects natural predispositions, biases, and inclinations, and is modeled as an L_1 action.

Of course, this is a stylized example, but human agents often behave in similar ways – acting without full awareness of the underlying causes of their decisions. This phenomenon has been extensively studied in behavioral decision-making and cognitive psychology, most notably in the work of Daniel Kahneman and collaborators [19]. In contrast, an L_2 action in a single-agent setting may involve

flipping a coin and allowing the coin toss to determine which machine to play. This process effectively averages over the agent's internal biases and corresponds to what is formally called the causal effect.

These L_1 and L_2 behaviors are fundamentally distinct from a counterfactual scenario, in which an agent intends to play machine $X = x$ but ends up playing $X = x'$. Formally, this is represented by the counterfactual quantity $P(Y_{X=x} \mid X = x')$. Note that if the agent naturally plays $X = x'$, this implies they were not initially inclined to play $X = x$ (see Fig. 2a for an illustration). This type of counterfactual evaluation was made possible by the introduction of counterfactual randomization [3], precisely to decouple the agent from its natural flow and enable the estimation of Q. [1] The theoretical limits of which counterfactuals can be physically inferred from the world have been recently characterized [34]. In positive cases, once the counterfactual randomization step is performed, the agent's intuition becomes a new type of information, which can then be conditioned upon.

2. Why are Causal Models essential if they can be converted into a matrix game with counterfactual actions?

A: Structural causal models (SCMs) are essential for constructing the payoff matrix (e.g., Fig. 3). In particular, determining the payoffs corresponding to L_1 and L_3 actions inherently relies on the underlying causal model. Once this matrix is computed, it may be viewed as a normal-form game. However, there are four critical phases where causal modeling plays a vital role:

- **Representation:** Traditional game theory lacks the concept of natural actions; it primarily deals with interventions, corresponding to L_2 actions in our causal framework. Once the existence of natural actions is acknowledged, the action space expands to include L_1, L_2, and L_3, enabling the construction of a richer payoff matrix.

  Example 1.1 illustrates this distinction more explicitly: two SCMs may yield identical payoffs and equilibria in the L_2 action space, yet diverge significantly when L_1 and L_3 actions are considered. This expanded matrix – and its associated equilibria – is difficult to recover without causal assumptions or an underlying structural model. For instance, suppose we observe repeated instances of the scenario in Ex. 3.1 and attempt to infer the payoffs for the actions (C, C), (C, D), (D, C), and (D, D) from observational data. We might obtain $(-1.4, -1.4)$, $(-8, 0)$, $(0, -8)$, and $(0, 0)$, respectively. However, if agents follow a randomized controlled trial (RCT) protocol, the corresponding payoffs could be $(-1, -1)$, $(-7, -0.5)$, $(-0.5, -7)$, and $(-1.9, -1.9)$. An underlying causal model with unobserved confounders can explain this discrepancy. The three layers of the causal hierarchy – observational, interventional, and counterfactual – were formally introduced in Bareinboim et al. [4].

- **Agency and Execution:** Causal modeling also addresses the practical question of how agents can implement these actions. While the payoff matrix in Fig. 3.1 encodes the outcomes, it does not specify the mechanisms by which those actions are executed. From prior results in causal decision theory, we know that L_1 actions correspond to passively observing the system, L_2 actions to standard interventions (e.g., RCTs), and L_3 actions require more advanced techniques, such as counterfactual randomization (e.g., ctf-RCT). Thus, causal modeling provides a bridge between abstract game-theoretic strategies and their realizability in practice. (In fact, the notion of counterfactual realizability – achieved through a specificd type of randomization – appears to exhaust what is physically implementable in the real world; for a more refined discussion, see Raghavan and Bareinboim [33].)

- **Solution Concept:** Our solution concept goes beyond merely computing a Nash equilibrium over an expanded action space. Causal modeling introduces a hierarchy of action spaces, allowing agents to commit to specific layers (e.g., L_1, L_2, $L_1 + L_2$, L_3) and ignore the others. This gives rise to a new "metagame," where the strategic choice involves selecting a layer of the PCH, and the equilibrium is computed within the corresponding subspace. This layered structure adds a new dimension to strategic reasoning and highlights the importance of causal structure in shaping the space of strategic possibilities.

- **Learning:** Normal-form representations overlook the structure linking agents' intuitions to their executed actions, whereas CNFGs can capture this relationship directly. In practice,

[1] There are intriguing connections here to the notion of free will (see Pearl's 2013 discussion, "The Curse of Free Will and the Paradox of Inevitable Regret").

agents may not observe the other agent's intuition when learning the payoff matrix. In such cases, this structural distinction becomes essential, and is explicitly exploited in Alg. 2.

3. Is the causal graph necessary? How does it help?

A: Causal graphs contain information about the structural dependencies between variables. For example, when the SCM is Markovian in a CNFG, the optimal actions or equilibrium strategies will always lie in L_2. However, when Markovianity does not hold, the optimal strategies may fall anywhere between L_1 and L_3. Without further knowledge of the parameters – either through prior knowledge or interaction with the system—it is impossible to determine which strategy is better.

4. How does this work relate to prior works?

A: Our goal in this paper is to develop a model that captures both instinctive and deliberate decision-making processes of the human mind. Structural Causal Models provide a principled framework for representing reality as closely as possible. Some prior works have attempted to reconcile causal models and game theory. However, due to differing goals, their models can be extremely restrictive, as discussed in detail in Appendix A.7.