# Counterfactual Rationality: A Causal Approach to Game Theory

**Aurghya Maiti** [1] **Prateek Jain** [1] **Elias Bareinboim** [1]

## 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 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 that has received significant attention due to its extensive application 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, each with their own objectives, preferences, and constraints, to make coherent and coordinated decisions within complex, dynamic environments. Agents may be individuals, autonomous systems, or organizations, with interactions that range from purely competitive settings to cooperative situations. 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 & Morgenstern, 1947) introduced the *expected utility theory*, providing a mathematical framework for *rational* decision-making, where agents select actions to maximize their expected utility. Since then, Game Theory (GT) has been the foundational framework for decision-making in MAS. Its models, such as Nash equilibrium (Nash Jr, 1950), cooperative game theory (Shapley, 1953), evolutionary game theory, and Bayesian games (Harsanyi, 1967) have been widely employed to study a range of scenarios where agents' decisions impact one another. Although rational decisions are grounded in systematic analysis and objective reasoning, human choices are often influenced by cognitive biases, emotions, social, and various unobserved factors that lead to seemingly irrational outcomes.

Irrational decisions may not be bad for all agents. In some settings, irrational or naive choices can result in better outcomes than rational ones, a phenomenon known as the *paradox of rationality* (Howard, 1971; Colman, 2003; Basu, 1994). Behavioral economics has tried to understand and model several forms of irrationalities, including loss aversion (Kahneman & Tversky, 1979), anchoring (Tversky & Kahneman, 1974), framing of choices (Kahneman & Tversky, 1984), social preferences (Fehr & Schmidt, 1999), and emotions (Loewenstein, 2003), to cite a few. (Kahneman, 2011) also proposed the *dual-process theory*, which suggests that humans have two systems for processing information and making decisions – a fast, automatic *System 1* and a slow, deliberate *System 2*. Although these methods model some aspects of our irrationalities and attempt to explain human decision-making, the general question of *when* and *how* players can strategically leverage such unobserved biases to their advantage in an MAS remains largely unexplored.

[1]Columbia University, New York, US. Correspondence to: Aurghya Maiti <aurghya@cs.columbia.edu>, Elias Bareinboim <eb@cs.columbia.edu>.

CausalAI Lab Technical Report, New York, United States. 

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

Table 1: $Y_1, Y_2$ as a function of $U_1, U_2, X_1, X_2$

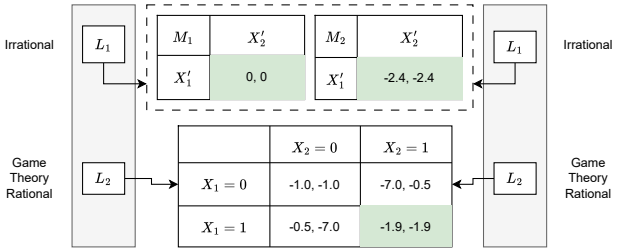


Figure 1: To be rational or not to be rational, that is the question.

In this work, we make a significant step towards addressing these issues by proposing a framework, rooted in the causal modeling proposed by (Pearl, 2009; Bareinboim et al., 2022). Research has shown that human decisions are often guided by causal structures (Tversky & Kahneman, 2015; Sloman & Hagmayer, 2006; Nichols & Danks, 2007), and actions can be thought of as interventions in a causal system (Hagmayer & Sloman, 2009). 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). Structural models have been successfully used in the context of decision-making, both for single-step bandit problems (Bareinboim et al., 2015; Zhang & Bareinboim, 2017) and for multi-step, more general RL problems (Lee & Bareinboim, 2020; Ruan et al., 2023), as surveyed in (Bareinboim et al., 2024). The advantage of such modeling is not only computational but more fundamental. Consider the example of *Greedy Casino*, introduced in (Bareinboim et al., 2015), where a randomized control trial (RCT) suggests that the expected payoff is higher than the realized payoff of players following their natural instincts (i.e., 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 exploited their natural biases, which, surprisingly, led to payoffs exceeding those based on the RCT.

In this paper, we build on these insights and model MAS through a causal lens, showing that existing game models may not capture some fundamental features of the decision-making process. This framework models agents' interactions within a system through the different layers of the Pearl Causal Hierarchy (Bareinboim et al., 2022). As a consequence, an agent will have the capability to act rationally (following Nash's prescription), irrationally, or as some mixture of both. We develop the notion of counterfactual rationality to formally understand if and how it is preferable for agents to act irrationally, when it is better not to. The next example illustrates why this task is nontrivial.

**Example 1.1** (Causal Prisoner's Dilemma (CPD))**.** *Two thieves are suspected in a crime, but due to insufficient evidence, they cannot be convicted outright. Now, they have a choice to make – either remain silent and cooperate ($C$) or betray the other by defecting ($D$). 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, the likelihood of 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 the 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. Mathematically, for prisoner $i$, their instinctive or natural choices are given by the function: $X_i \leftarrow f_X(R_i, U_i) = R_i \oplus U_i$, 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.*

*We consider 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 have a poor understanding of their circumstances ($R_1 = R_2 = 0$). In both cases, the situation is adversarial with probability $P(U_1 = 0) = P(U_2 = 0) = 0.6$. 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 detailed in Table 1. The values in the table indicate the prison sentences assigned to each prisoner based on their choices and circumstances. 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:*

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

*Notably, both scenarios* $M_1$ *and* $M_2$ *lead to the same Prisoner's Dilemma (PD) game, as shown in the* $2 \times 2$ *payoff table at the bottom of Fig. 1. 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. 1, 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: can we determine when it is beneficial to follow natural intuition and when it is better to override it, following Nash's prescription? In this paper, we explore the tension between rational and irrational behavior through a causal lens and derive from first principles how agents should deliberate and make decisions, thus resolving the so-called "paradox of rationality." Specifically, we outline our technical contributions.

1. We formalize a class of games that incorporate both rational and irrational behavior (Def. 2.10) and prove that this class is strictly more expressive than traditional Normal Form Games (Thm. 2.11).

2. We propose a new family of *counterfactual strategies* and establish the existence of equilibrium in the larger class of games (Thm. 3.5). We further demonstrate how such strategies can be better compared to 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 (Pearl, 2009). An SCM $M$ is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$, 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 $\{f_X : X \in \mathbf{X}\}$, that would normally determine their values. 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, refer to Appendix A.1 and (Bareinboim et al., 2022).

## 2. Causal Normal Form Games

In this section, we model the interaction of multiple agents in a system through the language of SCMs and the PCH layers. We first define a set of action nodes and reward signals for the agents in the system along with the SCM $M$.

**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 $M : \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, \mathbb{P} \rangle$ is an SCM and

- $N$ is the set of $n$ agents,
- $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ 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 $i \neq j$,
- $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ is the ordered set of reward signals, with $\mathbf{Y}_i \subseteq \mathbf{V}$ for all $i \in [n]$. □

A CMAS is essentially an SCM that contains nodes $\mathbf{X}$ that represent actions available to the $n$ agents in the system. Each agent has control over a distinct subset of action nodes; so, no two agents can act on the same variable. Also, the system contains a set of reward variables, $\mathbf{Y}$, which represent the feedback or payoff that 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:*

1. $\mathbf{U} = \{U_1, U_2, R_1, R_2\}$
2. $\mathbf{V} = \{X_1, X_2, Y_1, Y_2\}$
3. $X_i = R_i \oplus U_i$ *for* $i \in \{1, 2\}$*.* $Y_1, Y_2$ *as a function of* $U_1, U_2, X_1, X_2$ *are shown in Table 1.*
4. $P(U_i = 1) = 0.4, P(R_i = 0) = 1$ *for* $i \in \{1, 2\}$

*The elements of CMAS can now be defined as (i)* $N = \{1, 2\}$*, (ii)* $M = \mathcal{M}$*, (iii)* $\mathbf{X} = (\{X_1\}, \{X_2\})$ *and (iv)* $\mathbf{Y} = (\{Y_1\}, \{Y_2\})$*.*

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

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

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 or variable in the system. The $L_1$ action space is thus $\mathcal{A}^1 = \{a_0\}$ and $L_1$ 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, that is*

$$X_1 = R_1 \oplus U_1, \quad X_2 = R_2 \oplus U_2 \tag{2}$$

*The expected payoff when both the agents are following their natural intuition is given by*

$$\sum_{u_1,u_2,x_1,x_2} \mathbf{Y} \cdot P(u_1, u_2)P(x_1 \mid u_1)P(x_2 \mid u_2) \\ P(\mathbf{Y} \mid u_1, u_2, x_1, x_2) \approx (-2.4, -2.4) \tag{3}$$

*These and the other payoffs are shown in Fig. 1 (right).*

In a more traditional game theoretic sense, an agent can perform an intervention on the system. These interventions can be atomic interventions, where an agent sets the value of the action variable to a constant based on its context (Pearl, 2009), or soft interventions, where an agent samples their actions from a distribution (Correa & Bareinboim, 2020). Next, we define $L_2$ actions and the policy space.

**Definition 2.5** ($L_2$-action)**.** Given a CMAS $\langle N, M, \mathbf{X}, \mathbf{Y} \rangle$, $L_2$ 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 the action $do(\mathbf{x}_i)$ in the SCM $M$, then $\mathbf{X}_i$'s natural mechanism ($f_{\mathbf{X}_i}$) is replaced by

$$\mathbf{X}_i \leftarrow \mathbf{x}_i \tag{4}$$

The set of all such $L_2$ action will be denoted by $\mathcal{A}^2$. $L_2$ policy can be defined as a distribution over the actions in $\mathcal{A}^2$.

**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:*

$$X_1 \leftarrow 0, \quad X_2 \leftarrow 1, \tag{5}$$

*and $U_1, U_2, R_1, R_2$ are sampled from the distribution $P(\mathbf{U})$ as in Ex. 2.2. Similarly, $Y_1, Y_2$ are determined by Table 1. For instance, the expected payoff of the strategy $(do(X_1 = 0), do(X_2 = 1))$ will then be given by*

$$\sum_{u_1,u_2} \mathbf{Y} \cdot P(u_1, u_2)P(\mathbf{Y} \mid u_1, u_2, X_1 = 0, X_2 = 1) \tag{6}$$

$$\approx (-7.0, -0.5) \tag{7}$$

*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*

$$\sum_{u_1,u_2,x_2} \mathbf{Y} \cdot P(u_1, u_2)P(x_2 \mid u_2) \\ P(\mathbf{Y} \mid u_1, u_2, X_1 = 1, x_2) \approx (0, -8.9) \tag{8}$$

In many cases, an agent can interact with the environment through PCH's Layer 3 (Bareinboim et al., 2015; 2022; Raghavan & Bareinboim, 2025). This allows agents to incorporate certain counterfactuals into their decision-making. For example, in scenario $M_2$ of Ex. 1.1, following natural instinct led to a suboptimal outcome for both agents. However, if both agents had done the exact opposite of their instinctive choices, they could have achieved a payoff of $(0, 0)$. This ability to override instinct and strategically adjust behavior falls within the realm of Layer 3 of PCH. Before formally defining $L_3$ action, let $\mathbf{X}_i$ denote the action variable of the agent $i$, where its value is determined as a function $f_i$ of its observable and unobservable parents $Pa^+(\mathbf{X}_a)$.

**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 of an agent $i$ is when the value of $\mathbf{X}_i$ is determined by a mapping from natural intuition to action, denoted by $h : \mathcal{D}(\mathbf{X}_i) \to \mathcal{D}(\mathbf{X}_i)$. □

When an agent takes an $L_3$ action, they first note their natural instinct $\mathbf{X}'_i$ and then make the decision $\mathbf{X}_i$ as follows:

$$\mathbf{X}'_i \leftarrow f_i(Pa^+(\mathbf{X}_a)), \quad \mathbf{X}_i \leftarrow h_i(\mathbf{X}'_i) \tag{9}$$

In case $h(x) = x$, it is the same as the natural or $L_1$ action, and if $h(x)$ is constant for all $x$, then it is an intervention. In this light, we will often denote $a_0$ as $\mathbf{X} = \mathbf{X}'$, where $\mathbf{X}$ is the action variable and $\mathbf{X}'$ is the intuition.

**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 Player 1 and Player 2 are playing $g_1$ and $g_2$, respectively, then the variables are given by*

$$X'_i \leftarrow R_i \oplus U_i, \quad X_i \leftarrow g_i(X'_i) \tag{10}$$

*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 Table 1. For example, if $g_1(x) = 1 - x$ and $g_2(x) = 1 - x$, then the*

*expected payoffs are given by*

$$\sum_{u_1,u_2,x_1,x_2} \mathbf{Y} \cdot P(u_1,u_2)P(x_1 \mid u_1)P(x_2 \mid u_2)$$
$$P(\mathbf{Y} \mid u_1, u_2, X_1 = g_1(x_1), X_2 = g_2(x_2)) = (0,0) \tag{11}$$

Once the action spaces are defined, the policy space can be defined as a distribution over the action space. Let $\Delta(A)$ denotes the set of distribution over 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 $\mathcal{R}_a : \mathcal{D}(\mathbf{Y}_a) \to \mathbb{R}$ of an agent $a$ is a function from outcome $\mathbf{Y}_a$ to real numbers. □

In the CPD (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

- $\mathbb{M}$ is a CMAS $\langle N, M, \mathbf{X}, \mathbf{Y} \rangle$,
- $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_n)$ 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. □

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

**Theorem 2.11.** *Given a game in normal form, there exists two CNFGs $\mathcal{C}_1$ and $\mathcal{C}_2$ with expected $L_1$ payoffs $\mu_1$ and $\mu_2$ and Nash Equilibrium (NE) payoffs $\mu_{NE}$, such that*

$$\mu_2 \leq \mu_{NE} \leq \mu_1 \tag{12}$$

This result implies some important observations. First, the class of CNFGs is strictly larger (more expressive) than Normal Form Games (NFGs), meaning that some features of the decision-making process present in the real world cannot be expressed in terms of an NFG. Second, CNFGs provide a natural generalization of NFGs, aligning with the modern causal language, including the PCH. Third, in terms of practical decision-making, it is impossible to determine whether following natural intuition or deliberate actions is preferable without proper causal modeling. In the next section, we explore the properties of policy spaces and how to determine when to be irrational, rational, or counterfactual.

## 3. Causal Nash Equilibrium

In this section, we introduce counterfactual rationality and establish the Causal Nash Equilibrium 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 – whether to act instinctively ($L_1$), following a classical notion of rationality ($L_2$), like Nash, or engage in 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 moves 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. 2 shows the prisoner's payoff under these larger action space. If both prisoners follow their natural choices, playing $(X_1 = X_1', X_2 = X_2')$, their payoffs are $(-2.4, -2.4)$.*

*Now, suppose prisoner 1 starts thinking rationally, ignoring their natural instincts, which results in transition (a) in the figure. 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. 2. 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 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.*

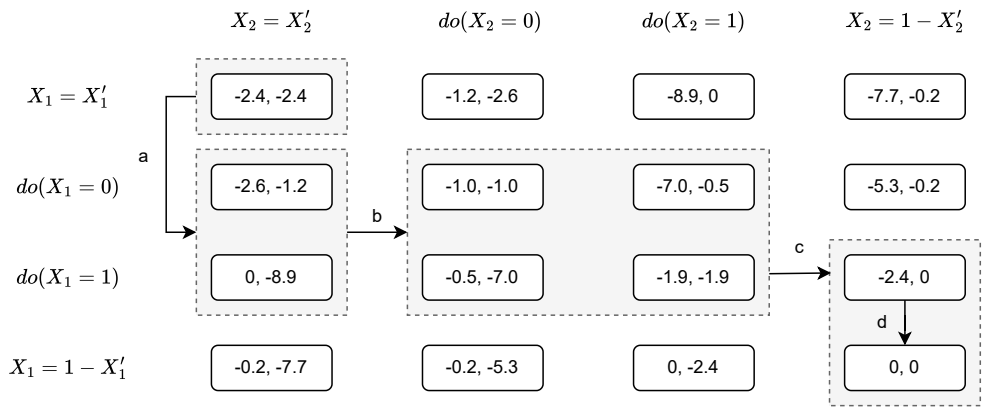


Figure 2: Change of Equilibrium with change of policies in Causal Prisoner's Dilemma.

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's always better to consider the full payoff table, since it provides the largest action space. To answer this, consider the example shown in Table 2. If Player 1's action space is limited to $L_1$ and $L_2$, 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 of CNFG)**.** Given a CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}_i \in \{\mathcal{A}_i^1, \mathcal{A}_i^2, \mathcal{A}_i^1 \cup \mathcal{A}_i^2, \mathcal{A}_i^3\}$. The PCH projection of $\Gamma$, denoted by $\Gamma(A_1, \ldots, A_n)$, is the subgame of $\Gamma$ where the action space of agent $i$ is constrained to a subset $A_i \subseteq \mathcal{A}_i$. □

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 key question now, is if we can 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 selection game, a meta-game, where agents choose which layer of PCH to operate at.

**Definition 3.3** (PCH Layer Selection Game)**.** Given a CNFG $\Gamma = \langle \mathbb{M}, \mathcal{A}, \mathcal{R} \rangle$, its PCH Layer Selection Game, or PCH-LSG, $L_\Gamma$ is the NFG with the same agents and

| Player 1 \ Player 2 | $X_2 = X_2'$ | $X_2 = 0$ | $X_2 = 1$ |
|---|---|---|---|
| $X_1 = X_1'$ | $-2, 2$ | $-2, -2$ | $-2, -2$ |
| $X_1 = 0$ | $0, 0$ | $-1.5, -1.5$ | $-1.5, -1.5$ |
| $X_1 = 1$ | $0, 0$ | $-1.5, -1.5$ | $-1.5, -1.5$ |
| $X_1 = 1 - X_1'$ | $2, -2$ | $-1, -1$ | $-1, -1$ |

Table 2: A Table showing that it is not always good for agents to jump to $L_3$ policy

- $A = A_1 \times \ldots, A_n$ is the action space where $\mathcal{A}_i \supseteq A_i \in \{\mathcal{A}_i^1, \mathcal{A}_i^2, \mathcal{A}_i^1 \cup \mathcal{A}_i^2, \mathcal{A}_i^3\}$,
- The utility function $u(A) = \texttt{NE}(\Gamma(A_1, \ldots, A_n))$,

where $\texttt{NE}(\Gamma(A_1, \ldots A_n))$ is a Nash Equilibrium payoff of the CNFG $\Gamma$ when actions spaces are restricted to $A_1, \ldots, A_n$. □

The PCH-LSG represents a higher-level decision process, where each cell in the payoff matrix corresponds to a PCH projection of $\Gamma$. The equilibrium of this game will determine the layer of reasoning at which agents should operate.

Let $s_i^*$ be the NE strategy of PCH-LSG. Let $\text{supp}(s_i^*)$ denotes the support of $s_i^*$, that is the action spaces, which has 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 $\Gamma$ that excludes $\mathcal{A}_i^j$. With all the components in place, we are ready to define the equilibrium under such causal strategies.

**Definition 3.4** (Causal Nash Equilibrium, or CNE)**.** Let $\Gamma$ be a CNFG and $L_\Gamma$ be its corresponding PCH-LSG, with NE strategy $s^*$. A strategy profile $\pi^*$ is called CNE if $\pi^*$ is the Nash Equilibrium of the PCH projection of $\Gamma$ with the

| Player 1 \ Player 2 | $\mathcal{A}^1$ | $\mathcal{A}^2$ | $\mathcal{A}^1 \cup \mathcal{A}^2$ |
|---|---|---|---|
| $\mathcal{A}^1$ | $-2, 2$ | $-2, -2$ | $-2, 2$ |
| $\mathcal{A}^2$ | $0, 0$ | $-1.5, -1.5$ | $0, 0$ |
| $\mathcal{A}^1 \cup \mathcal{A}^2$ | $0, 0$ | $-1.5, -1.5$ | $0, 0$ |
| $\mathcal{A}^3$ | $2, -2$ | $-1, -1$ | $-1, -1$ |

Table 3: PCH-LSG of the game $\Gamma$ presented in Table 2

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

Table 4: PCH-LSG for Causal Prisoner's Dilemma.

restricted action space $A^*$, defined as

$$A^* = A_1 \times \ldots \times A_n, \text{ where } A_i = \cup_{\mathcal{A} \in \text{supp}(s_i^*)} \mathcal{A} \quad (13)$$

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

If playing $L_2$ is the pure strategy Nash Equilibrium of the PCH-LSG $L_\Gamma$, then the CNE of $\Gamma$ in CGT and the NE of the normal-form game induced by $\Gamma$ in SGT coincides. Next, we look at how CNE compares with other action spaces.

**Theorem 3.6** (Dominance of causal strategies)**.** *Let $\Gamma$ be a CNFG with CNE payoff $\mu^*$ and $L_\Gamma$ be its PCH-LSG with NE strategy $s^*$. If $s^*$ is a pure strategy NE and $A_i^* = supp(s_i^*)$,*

$$\mu^* \geq \mathit{NE}(\Gamma(A_i, A_{-i}^*)) \quad (14)$$

*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 ensures that no agent can improve their payoff by switching to a different reasoning layer within the PCH framework, given that $L_\Gamma$ has a pure strategy NE. To illustrate this, consider Table 3 that represents the PCH-LSG for the game given in Table 2. The action spaces available to Player 1 and 2 in the CNFG $\Gamma$ are $(\mathcal{A}^3, \mathcal{A}^1 \cup \mathcal{A}^2)$. The full specification of $\Gamma$ is provided in Appendix D. Now, observe that $(\mathcal{A}^2, \mathcal{A}^1 \cup \mathcal{A}^2)$ is a pure strategy NE of the PCH-LSG. This means, that in the original game $\Gamma$, 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 policies. This leads to an interesting insight – CNE payoff of $\Gamma$ is thus $(0, 0)$. On the other hand, if both player had chosen $L_2$ policies, the NE payoff would be $(-1.5, -1.5)$, while choosing a $L_3$ policy for Player 1 would lead to a payoff of $(-1, -1)$.

**Algorithm 1** `Learn-CNE`

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

Now, consider the PCH-LSG in Table 4, which corresponds to the CPD described in Ex. 3.1. The pure strategy NE is $(\mathcal{A}^3, \mathcal{A}^3)$, implying that both players should adopt $L_3$ policies. This was indeed illustrated in Fig. 2.

## 4. Learning Causal Nash Equilibrium

In this section, we first introduce an algorithm that enables agents to infer the CNE when the payoff matrix is fully observable, as in typical game-theoretical settings. Then, we develop an algorithm to learn the payoff matrix with $L_3$ actions from observations and under partial observability.

Now, we consider the challenge of finding a CNE in a CNFG $\Gamma$, where action spaces available to the agents and their corresponding payoffs are available (as in SGT). For instance, if Player 1 has access to $L_3$ and Player 2 has access only to $L_2$ then both players are aware of the payoff of all possible action pairs generated by their respective action spaces. This assumption is equivalent to the PCH projections of $\Gamma$ being common knowledge. We introduce `Learn-CNE` (Alg. 1), which implements the ideas presented in Sec. 3. The algorithm first constructs PCH-LSG $L_\Gamma$ corresponding to $\Gamma$ using its PCH projections (Steps 3-4). Then, Step 5 computes its NE strategy. If an action space occurs with non-zero probability in the NE strategy, it is used for CNE, or else, we forget it (step 6). Step 7 computes the NE of the projection of $\Gamma$ with the restricted action space.

However, such dynamics of the game 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 action, but not their intuition. To this end, we propose an algorithm `Ctf-Nash-Learning` 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` (Bareinboim et al., 2024) and collect the dataset $(x_1', x_1, x_2, \mathbf{y})$. Note that for a fixed $(x_1', x_1, x_2)$, $\mathbf{y}$ is sampled from the mixture, $\sum_{x_2'} P(x_2' \mid x_1') P(\mathbf{y}_{x_1, x_2} \mid x_2', x_1')$.

Steps 3-4 identify the mean and the weights of the components of the mixture, which are essentially a permutation of $P(x_2' \mid x_1)$ and $E[\mathbf{Y}_{x_1, x_2} \mid x_1', x_2']$. For in-

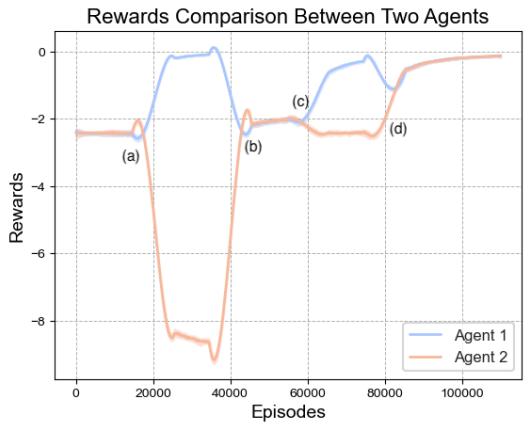


Figure 3: Change in payoffs of the players in Causal Prisoner's Dilemma move up the layers of PCH. Transitions (a), (b), (c) and (d) corresponds to the ones indicated in Fig. 2

stance, when we apply this to learn CNE in CPD, we see that $p_1(x'_1, x_1, x_2) \sim 0.6$ and $p_1(x'_1, x_1, x_2) \sim 0.4$ for all $(x'_1, x_1, x_2)$. Some examples of the means include $R_1(0,0,0) = (-1.5, -1.5)$ and $R_2(0,0,0) = (-1, -1)$, which corresponds to $E[\mathbf{Y}_{X_1=0,X_2=0} \mid X'_1 = 0, X'_2 = 0]$ and $E[\mathbf{Y}_{X_1=0,X_2=0} \mid X'_1 = 0, X'_2 = 1]$. These values can be consistently identified under certain technical assumptions (Appendix D). Step 5 handles the degenerate case when $\mathbf{Y}$ does not change with change in intuition. Step 6 defines the $L_3$ action spaces for the agents. In CPD, for agent 1, it is $\{f(x) = x, f(x) = 0, f(x) = 1, f(x) = 1 - x\}$ which corresponds to the actions $\{X_1 = X'_1, do(X_1 = 0), do(X_1 = 1), X_2 = X'_2\}$, and for agent 2, it is same $\{g(x) = x, g(x) = 0, g(x) = 1, g(x) = 1 - x\}$. The input Agent's functions 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 7 and the CNE strategy using `Learn-CNE`. The learnt probabilities, mean, and payoff matrix for CPD are shown in Appendix D.

**Theorem 4.1.** *Given a two player CNFG* $\Gamma = \langle \mathbb{M}, (\mathcal{A}^3_1, \mathcal{A}^3_2), \mathcal{R} \rangle$*, let* $s^*$ *be the NE strategy of the corresponding PCH-LSG* $L_\Gamma$ *and* $A_2 = \bigcup_{\mathcal{A} \in supp(s^*_2)} \mathcal{A}$*. If* $A_2 \in \{\mathcal{A}^2_2, \mathcal{A}^3_2\}$*, 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 (Tan, 1993), 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. 3. This is an experimental realization of the discussions presented in Ex. 1.1 and Fig. 2. Each 20,000 timesteps, one of the agents move up the layers of PCH: first, agent 1 (shown in blue) followed by agent 2 (orange).

**Algorithm 2** `Ctf-Nash-Learning`

1: **Input:** Dataset from `Ctf-RCT`: $(x'_1, x_1, x_2, \mathbf{y})$
2: **Output:** 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 $(y_1, y_2)$.
4: 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)
5: 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 $k$.
6: Define the action space for each player:

$$\mathcal{F}_1 = \{f : X'_1 \to X_1\}, \quad \mathcal{F}_2 = \{g : [k] \to X_2\}$$

7: 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 as:

$$\sum_{X'_1, i} P(X'_1) p_i(x'_1, f(x'_1), g(i)) R_i(x'_1, f(x'_1), g(i))$$

8: $(f^*, g^*) \leftarrow$ `Learn-CNE` on constructed payoff matrix without the action spaces $\mathcal{A}^1_2, \mathcal{A}^1_2 \cup \mathcal{A}^2_2$
9: **Return:** Strategy $f^*$.

## 5. Conclusions

In this work, we study the tension between rational and irrational decision-making, or the paradox of rationality, through a causal lens. In particular, we introduce the Causal Prisoner's Dilemma, where being rational is preferable in one setting and being irrational in another, while both settings induce the same game-theoretic solution. This presents a puzzle, as standard methods do not allow us to determine which choice is better. To formally understand this problem, we developed a causal framework capable of accounting for both rational and irrational behaviors, which was shown to be strictly more expressive than Normal Form Games (Thm. 2.11). Next, we introduced counterfactual strategies and established the properties of equilibrium under such strategies (Thm. 3.6). We further developed algorithms to learn Causal NE when the payoff matrix is common knowledge (`Learn-CNE`) and when it is unknown, but agents learn it through interactions (`Ctf-Nash-Learning`). We hope this work can help toward constructing more rational decision-making systems.

## 6. Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References


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

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. *On Pearl's Hierarchy and the Foundations of Causal Inference*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

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

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

Chan, L., Critch, A., and Dragan, A. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*, 2021.

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

Correa, J. and Bareinboim, E. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10093–10100, 2020.

Everitt, T., Carey, R., Langlois, E. D., Ortega, P. A., and Legg, S. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11487–11495, 2021.

Fehr, E. and Schmidt, K. M. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

Hagmayer, Y. and Sloman, S. A. Decision makers conceive of their choices as interventions. *Journal of experimental psychology: General*, 138(1):22, 2009.

Hammond, L., Fox, J., Everitt, T., Abate, A., and Wooldridge, M. Equilibrium refinements for multi-agent influence diagrams: theory and practice. *arXiv preprint arXiv:2102.05008*, 2021.

Hammond, L., Fox, J., Everitt, T., Carey, R., Abate, A., and Wooldridge, M. Reasoning about causality in games. *Artificial Intelligence*, 320:103919, 2023.

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

Howard, N. Paradoxes of rationality: theory of metagames and political behavior. *(No Title)*, 1971.

Howard, R. A., Oliver, R., and Smith, J. From influence to relevance to knowledge influences diagrams, belief nets and decision analysis. *Eds. RM Oliver and JQ Smith, John Wiley & Sons Ltd*, pp. 3–23, 1990.

Kahneman, D. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

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

Kahneman, D. and Tversky, A. Choices, values, and frames. *American psychologist*, 39(4):341, 1984.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Kearns, M., Littman, M. L., and Singh, S. Graphical models for game theory. UAI'01, pp. 253–260, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Koller, D. and Milch, B. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

Lauritzen, S. L. and Nilsson, D. Representing and solving decision problems with limited information. *Management Science*, 47(9):1235–1251, 2001.

Lee, S. and Bareinboim, E. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in neural information processing systems*, 33: 8565–8576, 2020.

Loewenstein, G. The role of affect in decision making. *Handbook of Affective Sciences/Oxford UniversityPress*, 2003.

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

Nichols, W. and Danks, D. Decision making using learned causal structures. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

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

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

Ruan, K., Zhang, J., Di, X., and Bareinboim, E. Causal imitation learning via inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Sloman, S. A. and Hagmayer, Y. The causal psycho-logic of choice. *Trends in cognitive sciences*, 10(9):407–412, 2006.

Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.

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

Tversky, A. and Kahneman, D. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

Tversky, A. and Kahneman, D. Causal schemas in judgments under uncertainty. In *Progress in social psychology*, pp. 49–72. Psychology Press, 2015.

Von Neumann, J. and Morgenstern, O. Theory of games and economic behavior, 2nd rev. 1947.

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

Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1778–1780, 2017.

# A. Preliminaries and Background

## A.1. Structural Causal Models and PCH

Structural Causal Models (Pearl, 2009; Bareinboim et al., 2024) is a general class of data-generating models found in the literature. It 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 heirarchical relation among them, known as Pearl Causal Hierarchy (PCH). Our presentation mostly follows (Bareinboim et al., 2022).

**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, (Bareinboim et al., 2022) ).

**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 forms of distributions that result from three types of interactions with the SCM.

**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 $\mathbf{Y} \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}) = \sum_{\mathbf{u} | \mathbf{Y}(\mathbf{u}) = \mathbf{y}} P(\mathbf{u}) \tag{15}$$

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 $\mathbf{Y} \subseteq \mathbf{X}$,

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}) = \sum_{\mathbf{u} | \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}} P(\mathbf{u}) \tag{16}$$

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

**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 $\mathbf{y}_{\mathbf{x}}, \ldots \mathbf{z}_{\mathbf{w}}$ for $\mathbf{Y}, \mathbf{Z}, \ldots, \mathbf{W}, \mathbf{X} \in \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots \mathbf{z}_{\mathbf{w}}) = \sum_{\mathbf{u} | \mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}, \ldots \mathbf{Z}_{\mathbf{w}}(\mathbf{u}) = \mathbf{z}} P(\mathbf{u}) \tag{17}$$

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

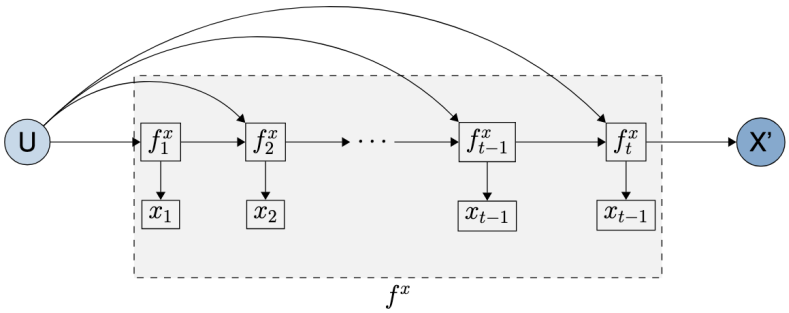


Figure 4: Illustration of decision flow $f_X$

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

Table 5: Payoff matrix for Prisoner's Dilemma

### A.2. Counterfactual Randomization

**Counterfactual Randomization.** (Bareinboim et al., 2015) introduces a novel form of randomization to interact through the Layer 3 of PCH. The core idea is to interrupt any reasoning agent just before they execute their choice, treat this choice as their intention, and then act. This procedure involve subtle issues, and we refer readers to Sec. 7 in (Bareinboim et al., 2024) 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. The same reference also proposed a procedure called `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.

### A.3. Normal Form Games

In this section, we provide definitions for some of the game theory concepts used in the paper.

**Definition A.6** (Normal Form Games)**.** A finite $n$-person normal form game is a tuple $\langle N, A, u \rangle$, where

- $N$ is the finite set of $n$ players indexed by $i$;
- $A = A_1 \times \ldots \times A_n$ where $A_i$ is the finite set of actions available to player $i$. Each vector $a = a_1 \times \ldots \times a_n$ is known as the action profile;
- $u = \{u_1, \ldots, u_n\}$ where $u_i : A \to \mathbb{R}$ is a real-valued utility (or payoff) function for player

A mixed strategy $S_i$ for an agent $i$ is a distribution over the elements of $A_i$. Next, we define best response and Nash Equilibrium.

**Definition A.7** (Best Response)**.** Player $i$'s best response to the strategy profile $s_{-i}$ is a mixed strategy $s_i^* \in S_i$ such that $u_i(s_i^*, s_{-i}) \geq u_i(s_i, s_{-i})$ for all strategies $s_i \in S_i$.

**Definition A.8** (Nash Equilibrium)**.** A strategy profiles $s = (s_1, \ldots, s_n)$ is a Nash equilibrium if, for all agents $i$, $s_i$ is a best response to $s_{-i}$.

### A.4. Paradox of Rationality

It has been observed several times in the Game Theory literature that irrationality can result in a better outcome than rational choices. One such example is the prisoner's dilemma. Both lying 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 (Colman, 2003). 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 (Tversky & Kahneman, 1981; Kahneman & Tversky, 1984; 2013) or through alternate forms of reasoning (Colman, 2003).

Consider the example of Travelers' Dilemma (Basu, 1994), 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, 1994) 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.5. 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 (Koller & Milch, 2003; Kearns et al., 2001). 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. (2001) 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 & Milch (2003) extended influence diagrams (Howard et al., 1990; Lauritzen & Nilsson, 2001) 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 related to how this relate to causality, including interventions and counterfactual reasoning.

The Structural Causal Influence Model by Everitt et al. (2021) connects causality with the influence diagrams literature (Howard et al., 1990; Lauritzen & Nilsson, 2001). 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 main challenges in typical causal settings.

Hammond et al. (2021) 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., 2023), 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.

The approach proposed by Chan et al. (2021) 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 randomization) 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.

## B. Proofs

### B.1. Proof of Theorem 2.11

Consider a Normal Form Game $\mathcal{G}$, with the action space $A = A_1 \times \dots A_n$ and the utility function $u = (u_1, \dots, u_n)$. Assume all utilites are finite. Suppose, we are looking at the payoff of agent $i$. Let $U = \sum_{a \in A} u_i(a)$ and the Nash Equilibrium payoff be $\mu^i_{NE}$. Consider a large number $M >> |U_i| + \mu^i_{NE}$. Now construct a CNFG $\Gamma_1$ with $\mathbf{U} = \{U_1, \dots, U_n\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_n\}$. The domain of $U_j$ is equal to $A_j$ for all $j$, and $P(U_j = a_j) = 1/|A_j|$, where $a_j \in A_j$. Also, for $j \neq i$, $Y_j(a, \mathbf{u}) = u_j(a)$. For agent $i$:

$$Y_i(a, \mathbf{u}) = u_i(a) + \mathbb{1}\{U_i = a_i\} \cdot M \cdot (|A_i| - 1) - \mathbb{1}\{U_i \neq a_i\} \cdot M \tag{18}$$

Construct $\Gamma_2$ same as $\Gamma_1$ except

$$Y_i(a, \mathbf{u}) = u_i(a) - \mathbb{1}\{U_i = a_i\} \cdot M \cdot (|A_i| - 1) + \mathbb{1}\{U_i \neq a_i\} \cdot M \tag{19}$$

Now, note that under $L_2$ action space, they induce the same NFG with the equilibrium payoff $\mu^i_{NE}$. However, for $\Gamma_1$ the $L_1$ payoff is $\mu^1_i > \mu^i_{NE}$ and in $\Gamma_2$, the $L_1$ payoff is $\mu^2_i < \mu^i_{NE}$. □

### B.2. Proof of Theorem 3.5

Consider the PCH-LSG $L_\Gamma$ corresponding to the CNFG $\Gamma$. Now since, $L_\Gamma$ is a NFG, a mixed strategy NE exists. Let this strategy be $s^*$. Consider the new action space $A^* = A^*_1 \times \dots \times A^*_n$, where $A^*_i = \text{supp}(s^*_i)$. This is a fixed policy space. The PCH projection of $\Gamma$ with $A^*$ is a subgame of $\Gamma$ 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 $L_\Gamma$, 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 $L_\Gamma$, 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. Observe, that since, $p_i$'s are arranged in descending order, their values will be identified consistently, that is if given $x'_1$, $X'_2 = 0$ occurs with a higher payoff, it will occur with a higher payoff, even when $X_2$ changes. Thus the algorithm can correctly distinguish between the different values of $X'_2$. However, the values cannot be identified, hence we will be able to learn only upto permutation of the values.

| 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 5: Payoff matrix for Prisoner's Dilemma with $L_2$ actions

| Player 1 \ Player 2 | $L_1$ | $X_2 = B$ | $X_2 = F$ |
|---|---|---|---|
| $L_1$ | $1.5, 1.5$ | $1, 0.5$ | $0.5, 1$ |
| $X_1 = B$ | $1, 0.5$ | $2, 1$ | $0, 0$ |
| $X_1 = F$ | $0.5, 1$ | $0, 0$ | $1, 2$ |

Figure 6: Payoff matrix for Prisoner's Dilemma with $L_1$ and $L_2$ actions

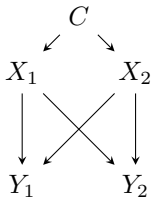


Figure 7: Causal Diagram for Battle of Sexes (Scenario 1)

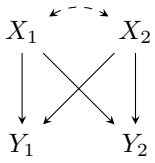


Figure 8: Causal Diagram for Battle of Sexes (Scenario 2)

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. Causal Games and Information

In this section, we show how the notion of information is orthogonal to the discussion and formalization within CNFG (Def. 2.10). This means that the framework discussed in the body can be immediately extended to include sources of information to the agent. First, we will look at the first an example with a correlation signal and how a CNFG will be different from it. Next, we will analyze an an example with a Bayesian Game to show how information can be incorporated in CNFG in a natural manner.

**Example C.1** (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 wants to go to a Football match. Their payoffs based on whether they are going to Ballet or Football are given in Fig. 5. The symmetric Nash equilibrium for this game is when both agents go to their preferred place two-third of the times and the joint payoff of such a strategy is* $(0.75, 0.75)$*. Now, consider two scenarios: the first is a classical example of correlated equilibrium.*

***Scenario 1 (correlated equilibrium):*** *Suppose the players have access to a coin and observe the outcome of the coin toss,* $\{H, T\}$*, following the strategy* $\{H \to B, T \to F\}$*. That is, if the coin shows heads, they will go to the ballet, and if it shows tails, they will go to watch the football match. Note that this is an equilibrium strategy, as neither of them has any incentive to deviate from it. The causal diagram for the game, along with the agents' policy, is shown in Fig. 7.*

***Scenario 2:*** *Now, suppose they are not able to coordinate using such a correlation signal. However, some external factors, such as advertisements about the success of a new ballet, may influence their intuitive or natural decisions to go to the ballet or the football match. Suppose these external factors are incorporated into an unobserved variable* $U$ *that influences the decisions* $X_1$ *and* $X_2$ *in the natural regime, or Layer 1 of PCH. Assume that the influence is such that either both of them decide to go to the ballet or both decide to go to the football match with equal probability. The agents, instead of having two actions, can now choose among three actions: follow their natural intuition (denoted by* $a_0$*), go to the ballet (*$B$*), or go to*

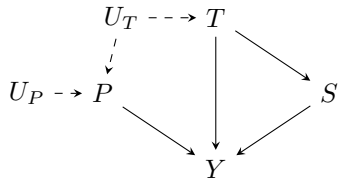


Figure 9: Causal Diagram for Sheriff's Dilemma with $L_1$ policy space.

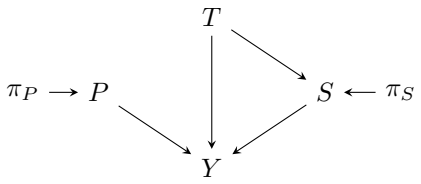


Figure 10: Causal Diagram for Sheriff's Dilemma with $L_2$ policy space.

*the football match ($F$). The causal diagram is shown in Fig. 8, and the payoffs are shown in Fig. 6. As we can see, following their natural intuition in this case is a symmetric Nash equilibrium.*

Scenarios 1 and 2 of Ex. C.1 demonstrate some important distinctions between the information structure studied in standard game theory and the concept of intuition in Layer 1 of PCH.

1. Intuition is not *trivially observed* as information.
2. Intuition, being an instance of the action variable, has the same domain as the action space. This is not true for an information structure.
3. In the example above, intuition did not require explicit coordination, unlike in Scenario 1, where the agents needed to decide what Heads and Tails represent.

Intuition and information are, in fact, distinct concepts, and both can be present in a multi-agent system without being at odds with each other. A richer representation of a game that allows both information for the agents and stochasticity of the rewards is a Bayesian game, as elaborated next.

**Definition C.2** (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 $\Theta_i$ 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 $i$

Causal Multi-Agent System can also contain state variables $\mathbf{S}$, other than the action variables $\mathbf{X}$ and the outcome variables $\mathbf{Y}$. Suppose each agent can condition their decision on information available to them. Suppose agent $i$ has access to $\mathbf{S}_i$. For example, now a hard intervention for agent $i$ would be a mapping $g$ from $D_{\mathbf{S}_i}$ to $D_{\mathbf{X}_i}$, that is $\mathbf{X}_i \leftarrow g(\mathbf{S}_i)$ instead of $\mathbf{X}_i \leftarrow \mathbf{x}_i$. This will be the $L_2$ action space with information. Similarly, we can define the $L_3$ action space for agent $i$ as the collection of mappings $h$ from $D_{\mathbf{S}_i} \times D_{\mathbf{X}_i}$ to $D_{\mathbf{X}_i}$, that is $\mathbf{X}_i \leftarrow h(\mathbf{X}'_i, \mathbf{S}_i)$. Thus information can be easily added to Causal Games. In terms of increasing information structure, the Nash Equilibrium (NE), Correlated Equilibrium (CE) and Bayesian Equilibrium (BE) are shown in Fig. 13. We have already seen causal extensions of two types of information structure corresponding to NE and CE. Next we provide an example for a Bayesian Game.

**Example C.3** (Causal Sheriff's Dilemma)**.** *A police officer faces an armed suspect, and they must simultaneously decide whether to shoot or not. 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 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.*

*The suspect's status as a criminal and the likelihood of shooting are probabilistic. In addition, unobserved factors might influence the officer's assessment of the suspect and the suspect's decision to shoot. For instance, a suspect's background might affect both their tendency to be a criminal and their behavior. A well-trained officer might be able to intuitively discern*

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

Figure 11: Payoff when suspect is criminal ($T=1$).

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

Figure 12: Payoff when suspect is civilian ($T=0$).

*these subtle cues and make a quick decision about whether to shoot. An untrained officer, on the other hand, may lack such abilities. This leads to unobserved confounding between the identity of the suspect and the officer's tendency to shoot.*

*The corresponding causal graph is shown in Fig. 9. In this graph, $T$ 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. Finally, $Y_1$ and $Y_2$ represent the utilities of the officer and the suspect, respectively.*

*Consider two scenarios, $M_1$ and $M_2$, which induce the same causal diagram. In $M_1$, the officers are well-trained, while in $M_2$, they are not. In both scenarios, let adverse backgrounds 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 are from an adverse background. This background may influence the suspect's behavior, which in turn can influence the officer's decision to shoot. In $M_1$, 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 11 and 12.*

*Now, suppose congress wants to recommend, through a new law, whether the officer should shoot or not. They compute the Bayesian Nash Equilibrium (BNE) of the game induced by the models and find that it is better if the officer does not shoot at all. Thus, the expected payoff of the officer under the BNE is given by:*

$$\mu_{BE} = -2 \cdot 0.1 = -0.2 \tag{20}$$

*However, if the law is not implemented, then in $M_1$ and $M_2$, the expected payoff of the policeman is respectively*

$$\mu_1 = -0.11, \quad \mu_2 = -0.99 \tag{21}$$

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

A table summarizing the increasingly refined information structure assumed as input to the different notions of equilibria is shown in Fig. 13.

## D. Additional Examples and Discussion

### D.1. SCM for Table 2

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 6.

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 $\phi_i$ 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

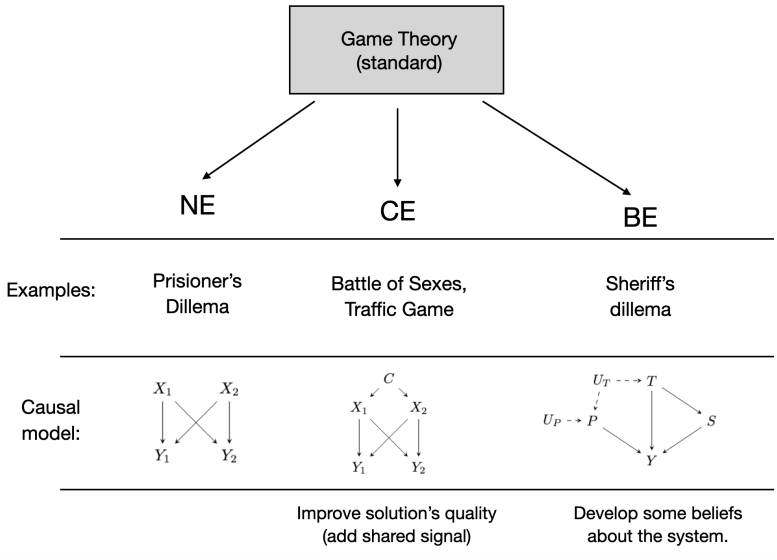


Figure 13: Increasing complexity of information structure

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

Table 6: $Y_1, Y_2$ as a function of $U_1, U_2, X_1, X_2$ for SCM in Table 2

their mixture upto a permutation of the $i$'s:

$$\sum_{i=1}^{k} p_i \phi_i(x_1', x_1, x_2) \tag{22}$$

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

Next, we show some 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** (Gaussiam Mixtures)**.** *(Yakowitz & Spragins, 1968) 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.

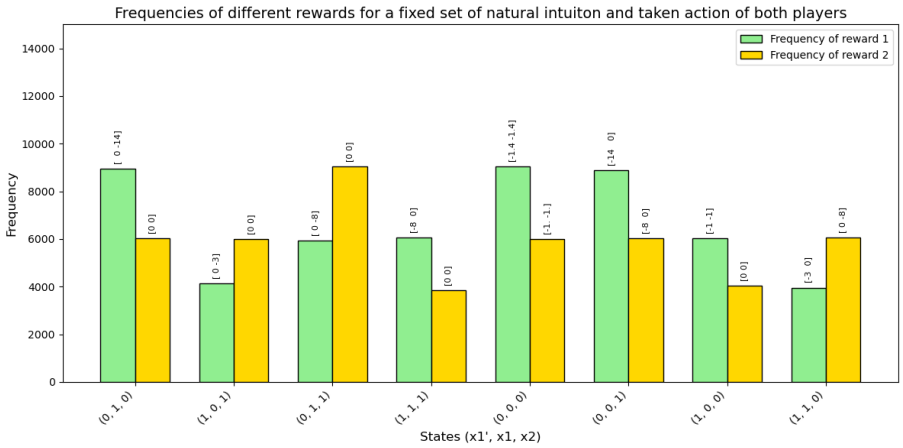


Figure 14: Frequencies of rewards observed for a particular tuple $(x_1', x_1, x_2')$

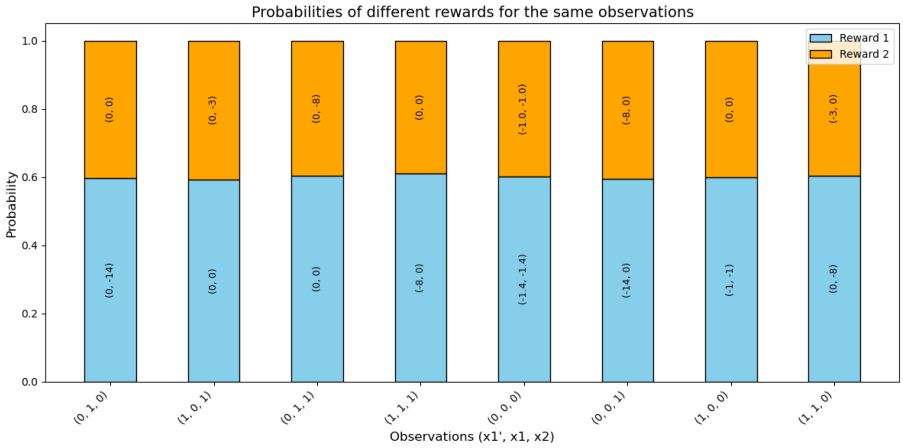


Figure 15: Probabilities of the rewards observed for a particular tuple $(x_1', x_1, x_2')$

### 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. 14. From this frequency table, we can compute the probabilities as shown in Fig. 15. The learned payoff matrix is shown in Table. 7. The code is available at `https://anonymous.4open.science/r/CGT-ICML/`.

Table 7: Payoff Matrix learned by Player 1 in Causal Prisoner's Dilemma

| | $X_2 = X_2'$ | $do(X_2 = 0)$ | $do(X_2 = 1)$ | $X_2 = 1 - X_2'$ |
|---|---|---|---|---|
| $X_1 = X_1'$ | $(-2.4428, -2.4450)$ | $(-1.2184, -2.6837)$ | $(-8.8923, 0.0000)$ | $(-7.6678, -0.2388)$ |
| $do(X_1 = 0)$ | $(-2.6828, -1.2345)$ | $(-0.9831, -0.9831)$ | $(-6.9321, -0.4901)$ | $(-5.2324, -0.2388)$ |
| $do(X_1 = 1)$ | $(0.0000, -8.8479)$ | $(-0.4753, -6.9509)$ | $(-1.9602, -1.8970)$ | $(-2.4354, 0.0000)$ |
| $X_1 = 1 - X_1'$ | $(-0.2400, -7.6374)$ | $(-0.2400, -5.2502)$ | $(0.0000, -2.3872)$ | $(0.0000, 0.0000)$ |