From inference to influence: applying causal game theory to complex security environments

Vonk, M.C.

Chapter 4

Causal Game Theory

While many causal concepts focus on isolated decision-making, game-theoretic extensions are germane to multi-agent contexts where strategic dynamics are central. These settings often involve actors whose choices depend not only on their own preferences but also on expectations about others' actions. To effectively integrate causal reasoning with game theory in modeling complex security environments, it is essential to understand both the foundational principles of game theory and their intersection with the previously introduced causal concepts.

This chapter systematically maps diverse game forms such as normal form, extensive form, and Bayesian games to their corresponding causal representations, synthesizing previously scattered connections within a probabilistic graphical model framework. Key concepts from (causal) game theory are elucidated, followed by an examination of the input required to operationalize such models. To bridge the gap between theoretical foundations and practical application, the chapter provides structured guidance for practitioners on selecting and applying these models in various security contexts. These concepts are further illustrated with examples derived from complex security environments, particularly those involving the deterrence of adversarial attacks.

In this way, RQ1.3 is addressed: *What methods exist for integrating causal reasoning with strategic decision-making in complex security environments, and how can they be applied?* This analysis not only advances theoretical understanding but also equips practitioners with the tools to model strategic interactions in a causally rigorous manner, fostering more robust decision-making in complex security contexts. Chapter 6 further examines one such tool to illustrate its practical relevance. The content of this chapter closely follows a paper [252].

4.1 Introduction

Game theory examines how rational decision-makers navigate strategic interactions, whether in competitive or cooperative settings. By modeling the decisions of interdependent agents, it provides a structured framework for analyzing incentives, strategy formation, and equilibrium outcomes. Recent research has aimed to combine the strengths of causal modeling and game theory [91, 227].

This chapter provides a structured framework that clarifies key concepts in game theory and its intersection with causality. By consistently adapting a practical example and referencing relevant research for implementation, it offers a concrete guide to navigating the complexities of integrating causal reasoning with game-theoretic modeling. This approach aims to bridge the divide between theory and practice, equipping practitioners with clearer implementation strategies and fostering closer collaboration between researchers and methodologists.

More specifically, the connection between causality and game theory in the context of PGMs [127, 236] is reviewed. The focus is on PGMs as they provide a structured

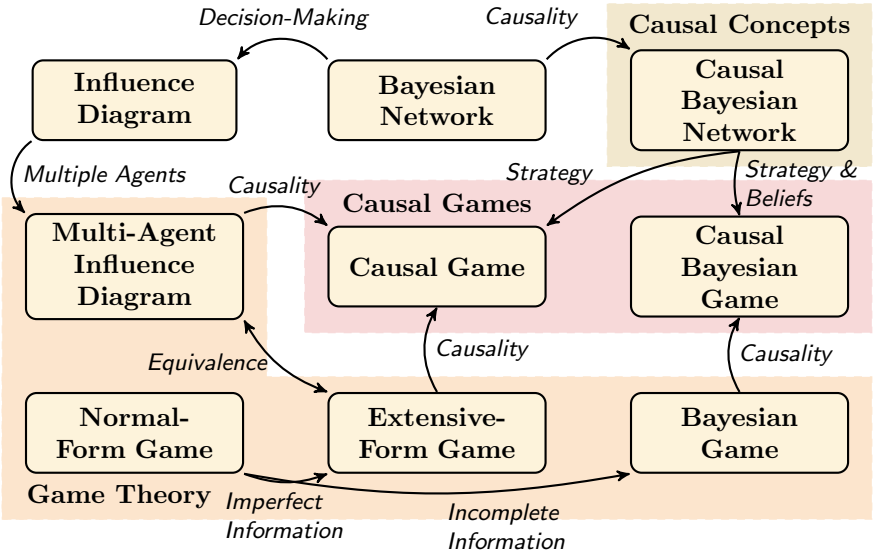


Figure 4.1: Scope of this chapter: the yellow blocks represent key concepts discussed within each domain: causality, game theory, and causal game theory. The concepts are grouped into categories that highlight their primary associations, indicated with background in different colors. Arrows indicate the possible extension or adaptation that allows for the transition from one concept to another.

approach to bridging the gap between theoretical advancements and practical implementation, particularly due to their intuitive representation of dependencies [183] and their capacity to unify diverse cognitive processes through a shared representational framework [58]. Through a detailed examination of their mathematical specifics, supported by illustrative examples from complex security environments, a structured framework that bridges theoretical understanding and practical implementation is provided. While exploring the mathematical details of various game-theoretic models and causal games might seem counterintuitive for practical purposes, it is necessary to articulate the distinctions between these models more precisely. This distinction not only helps practitioners select the most suitable model for their specific application but also clarifies the specific information required to work with these models effectively. Without delving into specifics, research that discusses techniques for eliciting the required information is pointed out. Finally, further considerations and insights are discussed to help surmount practical implementation challenges. The conceptual scope of this chapter is further illustrated in Figure 4.1, which categorizes the key concepts within the (causal) game-theoretical realm and their relation to previously discussed causal concepts.

Although doubts concerning the assumption of agent rationality [214, 83] have led to skepticism regarding the applicability of game-theoretic models [248, 204, 191], this thesis refrains from entering this philosophical debate.

The structure of this chapter is as follows. Section 4.2 introduces game-theoretic models, their solution concepts, and illustrative applications, along with a practical guide to their implementation. Section 4.3 builds on this foundation by extending game-theoretic models into the causal domain, presenting associated solution concepts, and discussing key considerations for their practical application. Conclusions and future research avenues are presented in Section 4.4.

4.2 Game Theory

To align causal concepts with the game-theoretical domain, this section focuses on game-theoretical components that have counterparts in causal reasoning. Therefore, three different types of games are considered: the normal-form game, the extensive-form game, and the Bayesian game. A normal-form game models strategic interactions in which players select actions simultaneously without knowledge of other players' choices, making it well-suited for competitive scenarios such as pricing strategies. An extensive-form game represents sequential decision-making, where players act in a

structured order, as in negotiations. A Bayesian game incorporates uncertainty, allowing players to make decisions based on private beliefs about unknown factors, making it particularly applicable to auctions.

The discussion begins by outlining formal game definitions and relevant solution concepts. Furthermore, the applicability of these game forms is explored by analyzing similar examples across different scenarios. Finally, the challenges associated with each form are addressed, and their practical utility is discussed. An overview of the game forms discussed, their concomitant characteristics, and required information for model implementation are summarized in Figure 4.4.

4.2.1 Normal-Form Game

First, the definition of a normal-form game and its associated solution concept is introduced. Then an example is provided.

Definition 4.1 (Normal-Form Game (NFG))**.** A *normal-form game* is a tuple $\Gamma = (M, \boldsymbol{A}, \boldsymbol{U})$ for which:

- $M = \{1, \ldots, m\}$ is a set of agents.
- $\boldsymbol{A} = \{A^1, \ldots, A^m\}$ is the set of action set, where A^i denotes the set of actions available to agent $i \in M$.
- $\boldsymbol{U} = \{u^1, \ldots, u^m\}$ is a set of utility functions where $u^i : \boldsymbol{A} \to \mathbb{R}$ is the payoff function for agent $i \in M$, representing the payoff that agent i receives.

Definition 4.2 (Nash Equilibrium)**.** A strategy profile $\hat{\sigma} = (\hat{\sigma}^1, \ldots, \hat{\sigma}^m)$ is a *Nash equilibrium* [167] if for every player $i \in \{1, \ldots, m\}$:

$$\hat{\sigma}^i \in \arg\max_{\sigma^i \in \boldsymbol{\Sigma}^i} u^i(\sigma^i, \hat{\sigma}^{-i}).$$

where $\boldsymbol{\Sigma}^i$ is player i's strategy space.

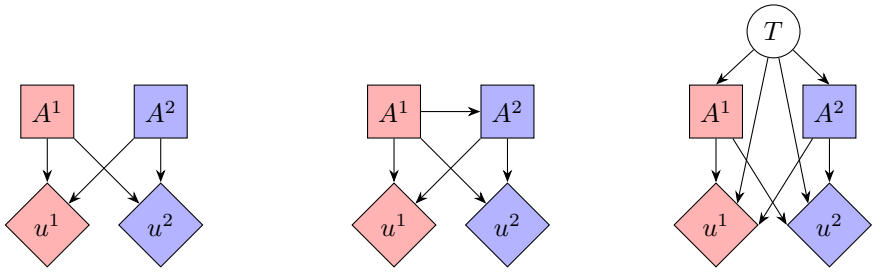
A Nash equilibrium represents a stable state of the game: given that all other players adhere to their equilibrium strategies $\hat{\sigma}^{-i}$, no player i can achieve higher utility by switching to any alternative strategy $\sigma^i \in \boldsymbol{\Sigma}^i$. Each player's equilibrium strategy is thus a best response to the equilibrium strategies of others, and no player has an incentive to unilaterally deviate. An illustration of the Nash equilibrium in a complex security environment follows.

Table 4.1: Utilities in Deterring Game (or Game of Chicken) in Normal-Form

		Adversary	
		a	$\neg a$
Deterrer	d	$(-1000, -1000)$	$(1, -1)$
	$\neg d$	$(-1, 1)$	$(0, 0)$

Example 6 (Deterring Game in Normal-Form)**.** Suppose a game models a strategic interaction between a deterring agent and its attacking adversary. The deterring agent threatens retaliation if attacked, but whether the agent is willing to follow through (d) or is bluffing ($\neg d$) is determined by action set A^1. Simultaneously, the adversary decides whether to attack (a) or not ($\neg a$), denoted by action set A^2. The game can be illustrated by Figure 4.2a and the game matrix is displayed in Table 4.1. The two pure Nash equilibria are ($\neg a$, d) and (a, $\neg d$).

Although the deterring agent acts first by issuing a threat of retaliation, the game-theoretical model does not consider the act of threatening as a strategic decision. Instead, the decision of interest is whether the deterring agent follows through on the threat. Since this decision does not occur after the attacker's move, both actions are effectively simultaneous and can be represented using a normal-form game. Nonetheless, it may very well be that acts do happen sequentially. In this case, a more refined game form is necessary, which is the extensive-form game.



(a) normal-form game **(b)** extensive-form game **(c)** Bayesian game

Figure 4.2: The relations of the normal-form game (a), extensive-form game (b), and Bayesian game (c). These relations correspond to Example 6 for (a), Examples 7 and 8 for (b), and Example 9 for (c). In the normal-form game, the deterring and attacking agents make independent decisions that determine their utilities. In contrast, the attacker's decisions are shaped by the deterrer's actions in the extensive-form game. Lastly, in the Bayesian game, the agents' decisions and utilities are influenced by their individual types and their beliefs about the opponent's type.

4.2.2 Extensive-Form Game

First, the definition of an extensive-form game is introduced, followed by an example and the relevance of a solution concept known as a subgame perfect equilibrium. The definition of Hammond et al. [91] is adopted.

Definition 4.3 (Extensive Form Game (EFG))**.** An *extensive-form game* is a tuple $\Gamma = (M, G, \boldsymbol{P}, \boldsymbol{A}, \lambda, \boldsymbol{I}, \boldsymbol{U})$ for which:

- $M = \{1, \ldots, m\}$ is a set of agents.
- $G = (\boldsymbol{V}, \boldsymbol{E})$ is a rooted tree, where the nodes $\boldsymbol{V}$ are partitioned into sets $\boldsymbol{V^0}, \boldsymbol{V^1}, \ldots, \boldsymbol{V^n}, \boldsymbol{T}$. In this case, $\boldsymbol{T}$ are the leaves or terminal nodes of G, $\boldsymbol{V^0}$ are chance nodes, and $\boldsymbol{V^i}$ are the decision nodes controlled by agent $i \in M$. The nodes are connected by edges $\boldsymbol{E}$.
- $\boldsymbol{P} = \{P_1, \ldots, P_{|\boldsymbol{V}^0|}\}$ represents a set of probability distributions P_j defined over the children of each chance node V_j^0, denoted as $\mathbf{ch}(V_j^0)$, for $j = 1, 2, \ldots, |\boldsymbol{V}^0|$.
- $\boldsymbol{A}$ represents the set of action sets, where $A_j^i \subseteq \boldsymbol{A}$ indicates the set of actions available at $V_j^i \in \boldsymbol{V^i}$.
- $\lambda : \boldsymbol{E} \to \boldsymbol{A}$ is a labelling function that assigns each edge (V_j^i, V_l^k) to an action $a \in A_j^i$.
- $\boldsymbol{I} = \{\boldsymbol{I}^1, \ldots, \boldsymbol{I}^m\}$ represents a collection of information sets, which partition the decision nodes controlled by agent i. Each information set $I_j^i \in \boldsymbol{I}^i$ is defined such that, for all $V_k^i, V_l^i \in I_j^i$, the available actions at these nodes are identical, i.e., $A_k^i = A_l^i$.
- $\boldsymbol{U} = \{u^1, \ldots, u^m\}$ is a set of utility functions where $u^i : \boldsymbol{T} \to \mathbb{R}$ is the payoff function for agent $i \in M$, representing the payoff that agent i receives.

Example 7 (Deterring Game in Extensive-Form)**.** Suppose a game models a strategic interaction between a deterring agent and its attacking adversary. This time, the deterring agent aims to deter by threatening retaliation denoted in action set A^1 in node V_1^1, which it is willing to follow through (d) or is bluffing ($\neg d$). Subsequentially, the adversary then acts A^2 to decide whether to attack (a) or not ($\neg a$) in nodes (V_1^2, V_2^2). Since the adversary does not have knowledge about the credibility of the deterrence effort, both nodes (V_1^2, V_2^2) are in the same information set, I_1^2. The dependencies of the game can be illustrated by Figure 4.2b and the game tree of Figure 4.3.

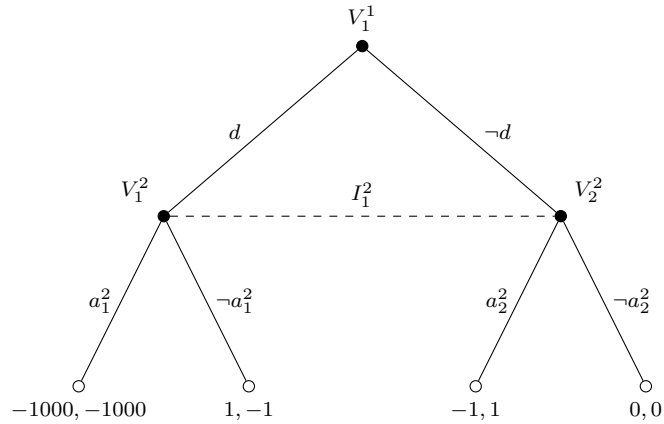


Figure 4.3: The figure illustrates a game tree of the deterring game in extensive-form.

Should the adversary have full information about the action of the deterring agent, a new game arises in which the adversary evaluates its actions based on this knowledge. Formally, this is known as a *subgame*.

Definition 4.4 (Subgame)**.** A *subgame* of an EFG is a game with a game tree $G' = (\boldsymbol{V}', \boldsymbol{E}')$ that is restricted to a node and its descendants such that any information set is either completely in the subgame or completely out the subgame.

Definition 4.5 (Subgame Perfect Equilibrium (SPE))**.** A *subgame perfect equilibrium* is a strategy profile σ such that for every subgame, σ constitutes a Nash equilibrium of the subgame.

Subgame perfect equilibria are relevant for the exclusion of non-credible threats, which are threats that a rational player has no incentive to carry out in later stages of the game. This is illustrated by the following example:

Example 8 (Deterring Subgame in Extensive-Form)**.** Consider a slightly modified version of Example 7 under conditions of perfect information, where the adversary recognizes the deterrer's commitment to enforce the threat, as shown in Figure 4.2b. The Nash equilibria are $((\neg a_1^2, \neg a_2^2), d), ((\neg a_1^2, a_2^2), d)$ and $((a_1^2, a_2^2), \neg d)$. In the subgame, the deterring agent anticipates that the adversary's threat to attack is not credible in the case of committed punishment, leaving $((\neg a_1^2, a_2^2), d)$ as the only subgame perfect equilibrium.

While *imperfect information* (uncertainty about the history of play) is captured directly through information sets, *incomplete information* (uncertainty about game

structure) requires modeling player types and beliefs. Bayesian games provide the framework for this.

4.2.3 Bayesian Game

The formal definition of a Bayesian game is first presented, followed by the introduction of its solution concept, the Bayesian Nash equilibrium. An example is then provided that extends the previous illustrations.

Although Bayesian games are sometimes introduced in terms of nature's states [174], the formulation in terms of players' types, as originally proposed [92, 76], is adopted here.

Definition 4.6 (Bayesian Game)**.** A *Bayesian game* is a tuple $\Gamma = \{M, \boldsymbol{A}, T, P, \boldsymbol{U}\}$, such that:

- $M = \{1, \ldots, m\}$ is the finite player set.
- $\boldsymbol{A}$ represents the set of action sets, where A^i is the action set of player i for $i \in M$.
- T^i is the finite type set of player i, and $t^i \in T^i$ its type. $T = (T^1, \ldots, T^m)$ is called the type profile tuple of Γ.
- $P : T \to [0, 1]$ is a probability distribution over T, referred to as the common prior. The belief of player i is denoted by

$$P(t^{-i} \mid t^i) = \frac{P(t^{-i}, t^i)}{P(t^i)} = \frac{P(t^{-i}, t^i)}{\sum_{t^{-i}} P(t^{-i}, t^i)},$$

  which describes player i's uncertainty about the other $m - 1$ players' possible types t^{-i}, given player i's type t^i, where $t^{-i} = (t^1, \ldots, t^{i-1}, t^{i+1}, \ldots, t^m)$ represents the tuple of the types of all the players except for player i.
- $\boldsymbol{U} = \{u^1, \ldots, u^m\}$ is a set of utility functions, where $u^i : T \times \boldsymbol{A} \to \mathbb{R}$ is the payoff function, which maps each action profile $\boldsymbol{a} \in \boldsymbol{A}$ to the pay-off of player i under each type profile $t^i \in T^i$.

Now that the structure of the game is contingent on the types of the players, the concept of a behavioral strategy can naturally be extended to account for these types: $\sigma^i(t^i) := \sigma^i(A^i \mid t^i)$ [22]. This allows the definition of a *Bayesian Nash equilibrium*.

Table 4.2: Utilities in Deterring Game in Bayesian Game

		Adversary Type t^1: $p = \frac{1}{4}$		Adversary Type t^2: $p = \frac{3}{4}$	
		a	$\neg a$	a	$\neg a$
Deterrer	d	$(-1000, -1000)$	$(1, -1)$	$(-1, 0)$	$(1, -1)$
	$\neg d$	$(-1, 1)$	$(0, 0)$	$(-1, 1)$	$(0, 0)$

Definition 4.7 (Bayesian Nash Equilibrium)**.** A strategy profile $\hat{\sigma} = (\hat{\sigma}^1, \ldots, \hat{\sigma}^m)$ is a *Bayesian Nash equilibrium* if for every player $i \in M$ and type $t^i \in T^i$:

$$\hat{\sigma}^i(t^i) \in \arg\max_{\sigma^i \in \mathbf{\Sigma}^i} P(t^{-i} \mid t^i) u^i(t^i, t^{-i}, \sigma^i, \hat{\sigma}^{-i}(t^{-i})).$$

Example 9 (Deterring Game in Bayesian Form)**.** Similar to the normal-form game, the actions of the deterring agent A^1 and its adversary A^2 are modeled independently. This time, the adversary can assume different types, where it is either protected against retaliation (t^2) or not (t^1). The relations of the game are illustrated by Figure 4.2c and the game matrices displayed in Table 4.2. The Bayesian Nash equilibria are $((\neg a, a), d)$ and $((a, a), \neg d)$. These are found by verifying that the deterrer maximizes expected utility given the probability distribution over adversary types ($p(t^1) = \frac{1}{4}$, $p(t^2) = \frac{3}{4}$), while each adversary type simultaneously maximizes their own payoff given their private information and the deterrer's strategy.

4.2.4 Practical Guide to Game Theory

Game theory offers a structured framework for organizing information on actors' decision-making processes [99]. Before applying game-theoretic concepts, a qualitative process should define the specific rules of the game for a given problem. This involves identifying stakeholders, outlining potential policy options, and establishing their interdependencies. Such an approach to framing policy problems is known as *metagame analysis* [107].

Practitioners should select the game type that best fits the policy problem's structure. Simultaneous decisions suit normal-form and Bayesian games,[1] whereas sequential decisions align with extensive-form games. Additionally, extensive-form games may involve imperfect information, while Bayesian games feature incomplete information. Although selecting a specific game type is useful, some scholars argue that a robust analytical approach benefits from modeling pluralism rather than strict unifor-

[1] Bayesian games can also be extended to sequential decision-making [174].

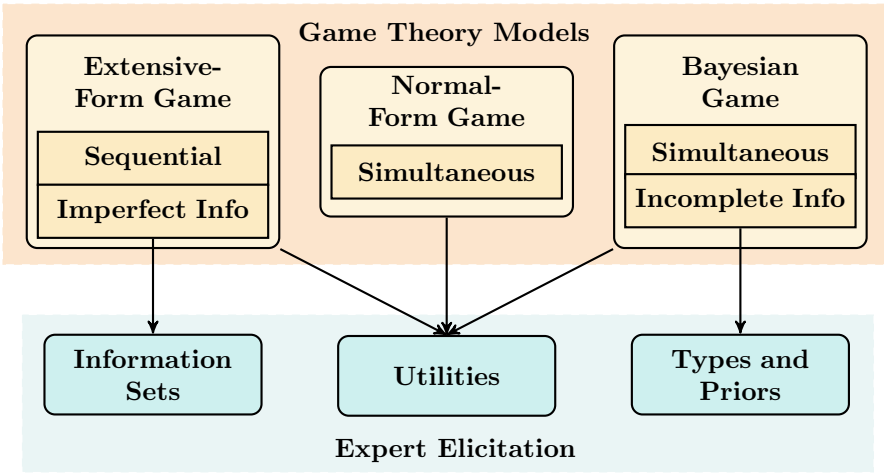


Figure 4.4: Elicitation requirements for different game-theoretic models: while the yellow blocks correspond to the different games along with their associated properties, the light green blocks indicate the different pieces of information that are required to be elicited. The arrows indicate what information is relevant to each game.

mity [6], advocating for the use of diverse game-theoretic frameworks where necessary while ensuring that variation remains focused on essential structural distinctions.

As the mathematical dissection of different games shows, each game requires specific information to be gathered before implementation. All games need the elicitation of utilities or preferences, with extensive-form and Bayesian games requiring more detailed utility structures. While utility elicitation is a demanding task, it has been well-studied and the reader is referred to the work by Wakker and Deneffe [254] for details about elicitation methods. If preference relations are uncertain, alternative methods can still provide insights into the stability of certain solution concepts [143].

In extensive-form games, information sets—clarifying who knows what at each decision point—need to be extracted. Additionally, chance nodes require probability distributions for the uncertainties they introduce. For Bayesian games, prior and posterior probabilities of types must also be elicited, a challenging process with guidelines written by Mikkola et al. [161]. The characteristics of the different games, along with the required elicitation information can be observed in Figure 4.4.

To bridge the gap between theory and application in game theory, a growing body of work focuses on deriving game-theoretic models from simulation data to enhance empirical validity. A comprehensive survey of recent advances in this area is provided by Wellman et al. [257].

4.3 Causal Game Theory

In this section, the intersection of game theory and causality is explored by integrating causal concepts from the previous chapter into the decision-making framework. Specifically, the Bayesian networks are extended to influence diagrams [108], distinguishing between purely probabilistic structures and decision-theoretic elements. This framework is further generalized to multi-agent settings, leading to multi-agent influence diagrams [128, 91], which form the foundational structure of causal games. To account for uncertainty in the causal structure of a game, the approach of Gonzalez Soto et al. [227] is adapted, introducing the causal Bayesian game. Finally, key considerations for the practical implementation of these models are discussed.

To establish a foundation for causal game theory, previous examples are reformulated through a causal lens. This provides a basis for extending causal reasoning into the strategic reasoning domain.

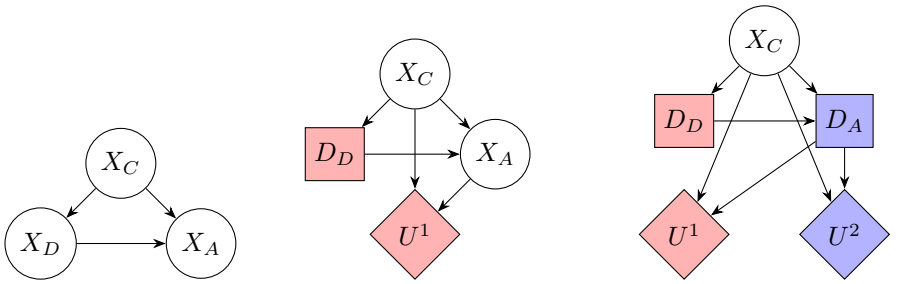
Example 10 (Deterring Relations in Causal Form)**.** Suppose there is observational data on the explicitness of deterrence messages X_D, which can either be explicit (d) or vague ($\neg d$). The goal of an explicit message is to dissuade the adversary from committing to an aggressive operation X_A ($X_D \to X_A$). However, both the explicitness of the deterrence message and the adversary's decision to attack are shaped by the deterring agent's military and strategic capabilities X_C, which can be strong (c) or weak ($\neg c$). These capabilities influence the explicitness of the message $X_C \to X_D$, but also directly affect the adversary's decision to commit to an aggressive operation $X_C \to X_A$. The causal relations are displayed in Figure 4.5a.

To compute the causal effect of explicit deterrence messaging on successful dissuasion, the covariate, the deterrer's capabilities, should be adjusted for. Therefore, the *do*-operator can be deployed, and the truncated factorization can be utilized:

$$P(X_A \mid do(X_D = d)) = \sum_{x_C \in \{c, \neg c\}} P(X_A \mid X_D = d, X_C = x_C) P(X_C = x_C).$$

4.3.1 Influence Diagram

An influence diagram extends a Bayesian network to the decision-making realm by dissecting the nodes into chance nodes, utility nodes, and decision nodes. More formally:



(a) Bayesian Network **(b)** Influence Diagram **(c)** MAID

Figure 4.5: The relations of the Bayesian network (a), influence diagram (b), and multi-agent influence diagram (c) of Example 10, 11 and 12, respectively. The Bayesian network does not distinguish between chance, decision and utility nodes as do the influence diagrams. In addition, the multi-agent influence diagram models the decision of the adversary also strategically.

Definition 4.8 (Influence Diagram)**.** An *influence diagram* contains a graphical structure $G = (\boldsymbol{V}, \boldsymbol{E})$ where $\boldsymbol{V}$ is separated into decision nodes $\boldsymbol{D}$, chance nodes $\boldsymbol{X}$ and utility nodes $\boldsymbol{U}$. Whereas the conditional probability distributions of $\boldsymbol{X}$ and $\boldsymbol{U}$ are known, any *decision rule* $\sigma(D)$ with $D \in \boldsymbol{D}$ corresponds to a conditional probability distribution over the decisions and hence all decision rules constitute the full joint probability distribution $P(\boldsymbol{V})$.

Example 11 (Deterring Relations as Influence Diagram)**.** A refinement of Example 10 involves separating the chance node of the military and strategic capabilities X_C from the decision node of deterrence messaging D_C. In this framework, the aggressor's decision to conduct an aggressive operation X_A can be modeled as another chance node, which is followed by a final utility node of the deterring agent U^1. These relations are illustrated by Figure 4.5b.

Influence diagrams incorporate decision-making elements into Bayesian networks,[2] but they only model the decision-making of a single agent, meaning there is no strategic interaction between agents. Strategic considerations emerge only when multiple agents make decisions in response to each other, as seen in multi-agent influence diagrams.

4.3.2 Multi-Agent Influence Diagram and Causal Game

First, multi-agent influence diagrams [128] are introduced, followed by the definition of a causal game [91]. The latter concept will then be illustrated with an example.

[2] In a similar way, causal Bayesian networks can be extended to *causal influence diagrams* [75].

Definition 4.9 (Multi-Agent Influence Diagram (MAID))**.** A *multi-agent influence diagram* contains a graphical structure $G = (\boldsymbol{V}, \boldsymbol{E})$ and a set of agents $M = \{1, \ldots, m\}$. Furthermore, the nodes $\boldsymbol{V}$ are separated in decision nodes $\boldsymbol{D} = \cup_{i \in M} \boldsymbol{D}^i$, chance nodes $\boldsymbol{X}$ and utility nodes $\boldsymbol{U} = \cup_{i \in M} \boldsymbol{U}^i$. Each strategy $\sigma^i(D^i)$ with $D^i \in \boldsymbol{D^i}$ defines a conditional probability distribution over a decision node. Consequently, given conditional probability distributions of $\boldsymbol{X}$ and $\boldsymbol{U}$, a complete strategy profile σ constitutes the full joint probability distribution $P(\boldsymbol{V})$.[3]

The causal game Γ associated with a MAID can be seen as a more abstract form of a MAID, where the parameters of the decision variables are yet to be defined.

Definition 4.10 (Causal Game)**.** A *causal game* Γ is a MAID such that for any chosen strategy profile σ, the induced joint probability distribution $P^\sigma(\boldsymbol{V})$ corresponds to a causal Bayesian network.

Similarly to extensive-form games, the Nash equilibrium of causal games can be defined in terms of the strategy profiles:

Definition 4.11 (Nash Equilibrium)**.** A strategy profile $\hat{\sigma} = (\hat{\sigma}^1, \ldots, \hat{\sigma}^m)$ is a *Nash equilibrium* if for every player $i \in \{1, \ldots, m\}$:

$$\hat{\sigma}^i \in \arg\max_{\sigma^i \in \boldsymbol{\Sigma}^i} \sum_{U \in \boldsymbol{U}^i} \mathbb{E}_{[\sigma^i, \hat{\sigma}^{-i}]}[U].$$

Multi-agent influence diagrams are powerful models, because they allow for the calculation of equilibria as well as the computation of policy interventions.

Example 12 (Multi-Agent Influence Diagram Deterring Game)**.** A multi-agent influence diagram can be derived from Example 11 when the adversary's decision to conduct an aggressive operation a or refrain from one $\neg a$ is modeled as a decision node of another agent D_A. This decision node is influenced by the deterring agent's decision node, D_D, which represents whether the deterrence messages are explicit (d) or vague ($\neg d$). Alongside the deterring agent's utility node U^1, the adversary also

Figure 4.6: Utilities of the Deterring Agent (left) and Attacking Agent (right)

U^1	$X_C = c$	$X_C = \neg c$
$D_A = a$	0	-1
$D_A = \neg a$	1	1

U^2	$X_C = c$	$X_C = \neg c$
$D_A = a$	-1000	1
$D_A = \neg a$	-1	-1

[3]Since EFGs and MAIDs are proven to be equivalent [91], σ is used again for a strategy profile.

possesses a utility node U^2. Both utility nodes depend on the attacker's decision D_A and the deterrer's capabilities X_C, which can be either strong (c) or weak ($\neg c$), with an equal probability distribution. The relations are displayed in Figure 4.5c and the utilities are further specified in Figure 4.6.

The game has eight Nash equilibria, one of which is equilibrium $\hat{\sigma}$ where the deterring agent issues an explicit deterrence message when possessing strong capabilities and a vague one otherwise. In this scenario, the adversary chooses not to attack if and only if the deterring agent demonstrates strong capabilities regardless of the deterrence message. The equilibrium $\hat{\sigma}$ also happens to be a subgame perfect equilibrium.[4] Within this equilibrium, the expected utility of the deterring agent for sending out an explicit message is $\mathbb{E}_{[\hat{\sigma}]}[U^1 \mid D_D = d] = 1$. Intervention effects can also be assessed in this equilibrium; for instance, if allied agents force an explicit deterrence message regardless of its capability, the utility becomes $\mathbb{E}_{[\hat{\sigma}]}[U^1 \mid do(D_D = d)] = 0$.

This example is a *post-policy* intervention as the results are computed after a strategy profile from the Nash equilibrium has been chosen. In contrast, *pre-policy* interventions allow agents to adjust their strategy profile after an intervention, which requires a more refined notion of a MAID [91]. The introduction of these notions, while relevant for strategic reasoning, is considered beyond the scope of the chapter as they do not bring additional implications for their practical implementation.

4.3.3 Causal Bayesian Games

The notion of a causal Bayesian game [227] was developed in order to allow for uncertainty about a graphical structure controlling an environment in which agents are located. The definition and notation of Soto et al. [227] is refined to align with the previously introduced causal games while ensuring consistency with earlier introduced Bayesian games. Following the Bayesian game in Section 4.2.3, the types correspond to distinct MAIDs within the family of causal graphical structures $\mathcal{G}$. Moreover, as in Bayesian games, players act independently, implying that a subset of MAIDs is considered where no direct causal paths exist between the decision nodes of different agents. Unlike the earlier introduced Bayesian game, players agree on the common state space $\mathcal{G}$ of possible graphical models but have private beliefs about the probability of these states $\mu_i(\mathcal{G})$. Naturally, interventions on decision nodes induce variations in payoffs across different graphical models $G \in \mathcal{G}$.

[4] Although the subgame perfect equilibrium of Definition 4.5 naturally extends to causal games, the notion of subgames in causal games is much richer than in EFGs. The reader is referred to the work of Hammond et al. [91] for details on this.

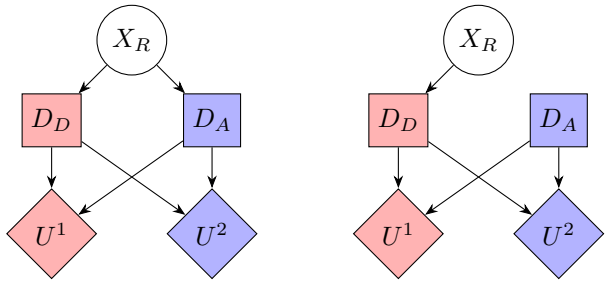
Definition 4.12 (Causal Bayesian Game)**.** Consider a family of different causal structures $G \in \mathcal{G}$ where no direct paths exist between the decision nodes of different agents. Each agent $i \in \{1, \dots, m\}$ has a private belief about the probability of these causal structures $\mu_i(\mathcal{G})$ and a higher-order belief $\mu_i(\mu_{-i}(\mathcal{G}))$, which reflect uncertainty over other $m-1$ players' beliefs. Each strategy $\sigma^i(G) = \sigma^i(D^i \mid G)$ defines a conditional probability distribution over a decision node $D^i \in \boldsymbol{D^i}$ conditional on the belief $\mu_i(G)$ of that graphical model $G \in \mathcal{G}$ for agent $i \in \{1, \dots, m\}$.[5] A *causal Bayesian game* Γ is a MAID such that for any chosen strategy profile σ, the induced joint probability distribution $P^{\sigma}(\boldsymbol{V})$ corresponds to a causal Bayesian network.

Note that the causal Bayesian network is not only induced by the strategies for each player but by the strategies of the players conditioned on the same graphical model. Naturally, these considerations are also reflected in the Bayesian Nash equilibrium.

Definition 4.13 (Bayesian Nash Equilibrium)**.** A strategy profile $\hat{\sigma} = (\hat{\sigma}^1, \dots, \hat{\sigma}^m)$ is a *Bayesian Nash equilibrium* if for every player $i \in \{1, \dots, m\}$ and graphical structure $G \in \mathcal{G}$:

$$\hat{\sigma}^i(G) \in \arg\max_{\sigma^i \in \boldsymbol{\Sigma}^i} \mu_i(\mu_{-i}(G)) \sum_{U \in \boldsymbol{U}^i} \mathbb{E}_{[\sigma^i(G), \hat{\sigma}^{-i}(G)]}[U].$$

Example 13 (Causal Bayesian Deterring Game)**.** Suppose two agents have their private beliefs about the causal structure of the game they are playing. In both games, players make decisions to defend D_D and attack D_A independently. While the utility of



(a) MAID G_1 **(b)** MAID G_2

Figure 4.7: The different causal structures G_1 and G_2 of the causal Bayesian game as in Example 13.

[5] Although the *do*-operator is applied in the original paper, it is relaxed here to allow for the implementation of more dynamic strategies.

Figure 4.8: Utilities of the Deterring Agent (left) and Attacking Agent (right) for Causal Structure G_1 (top) and G_2 (bottom)

$U^1(G_1)$	$D_A = a$	$D_A = \neg a$
$D_D = d$	-1000	1
$D_D = \neg d$	-1	0

$U^2(G_1)$	$D_A = a$	$D_A = \neg a$
$D_D = d$	-1000	-1
$D_D = \neg d$	1	0

$U^1(G_2)$	$D_A = a$	$D_A = \neg a$
$D_D = d$	-1	1
$D_D = \neg d$	-1	0

$U^2(G_2)$	$D_A = a$	$D_A = \neg a$
$D_D = d$	0	-1
$D_D = \neg d$	1	0

both agents is the result of both agents' actions, the attacking agent's decision can also be shaped by the defending agent's capability to retaliate X_R. Figure 4.7 illustrates the two different causal structures under consideration. While the attacking agent does not have access to the defending agent's retaliation capacity nor does he think the defending agent thinks he has ($\mu_A(G_2) = 1$ and $\mu_A(\mu_D(G_1)) = 1$), the defending agent considers a scenario where the attacking has access to his retaliation capacity with equal probability: $\mu_D(G_1) = \mu_D(G_2) = \mu_D(\mu_A(G_1)) = \mu_D(\mu_A(G_2)) = \frac{1}{2}$. Taking into account the utilities for the different structures indicated by Table 4.8, the Bayesian Nash equilibria are $((\neg a,a), d)$ and $((a,a), \neg d)$.

While this game and associated Bayesian Nash equilibrium is similar to Example 9, it is important to emphasize that uncertainty about the causal structure in this example only gives rise to different pay-offs. When alternative causal structures yield distinct payoff configurations and more sophisticated higher-order beliefs are involved, significant complications may arise.

4.3.4 Practical Guide to Causal Game Theory

Analogous to selecting a game-theoretic model, the choice of a causal game-theoretic model should consider whether agents possess private information regarding the causal structure. As the mathematical dissection discerns different types of nodes within causal games, practitioners must also clearly differentiate between decisions, chance events, and utilities. This distinction can be subtle, as illustrated in Example 12: a deterrer's capability, often modeled as a chance node, may not truly qualify as such if the agent has the option to enhance their capabilities. Consequently, this classification requires careful consideration, thoughtfully aligned with the specific research question at hand.

Unlike the standard Bayesian network in Example 10, which consists solely of

chance nodes, causal games with decision nodes do not require the elicitation of conditional probability distributions for those decisions, as they are being solved in response to the adversary. This game-theoretic aspect in causal games thus reduces some of the elicitation burden. The remaining conditional distribution of the chance nodes and the specifications in the utility nodes can be extracted via the elicitation methods introduced in Section 3.4.4 and Section 4.2.4, respectively.

While extracting higher-order beliefs in addition to uncertainty over the nature of graphical structure may appear highly complex, the methods for eliciting prior and posterior probabilities outlined in Section 4.2.4 remain applicable [161]. However, the increased complexity associated with calculating the relevant solution concepts across different causal structures may impede practical implementation.

A summary of the causal game and the causal Bayesian game along with required information for implementation is given in Figure 4.9.

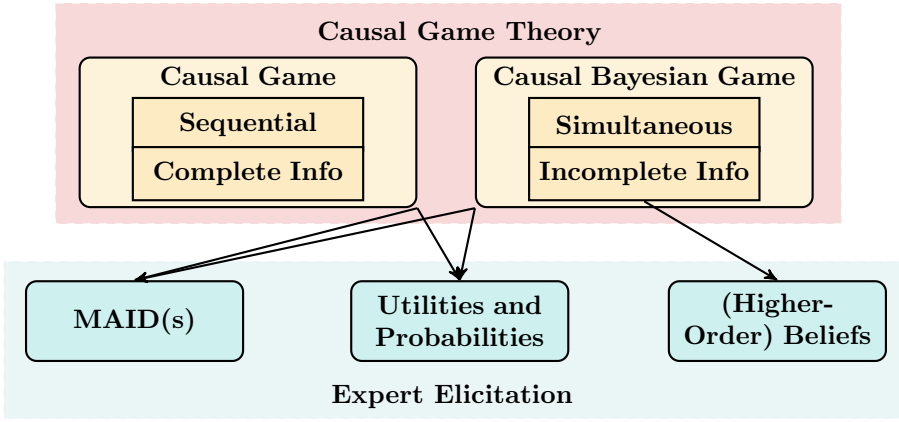


Figure 4.9: Elicitation requirements for causal games and causal Bayesian games and associated characteristics: a causal game necessitates the elicitation of a MAID, utilities, and conditional probabilities. In contrast, a causal Bayesian game further requires the elicitation of multiple MAIDs along with beliefs about the probabilities over these graphs and higher-order beliefs about other players' beliefs.

4.4 Conclusion and Future Work

This chapter has examined game-theoretical models and their integration with previously introduced causal concepts within the framework of probabilistic graphical models. The distinctions between various model types and the input required for their

implementation have been clarified. Practical guidelines, detailed examples, and considerations for effective application have also been provided. Future research could focus on developing integrated tools that unify model elicitation and implementation, streamlining the application of causal game-theoretical models [120, 29].