Causal Foundations of Collective Agency

Frederik Hytting Jørgensen FREDERIK.HYTTING@MATH.KU.DK
Copenhagen Causality Lab and Pioneer Centre for AI,
University of Copenhagen, Denmark

Sebastian Weichwald SWEICHWALD@MATH.KU.DK
Copenhagen Causality Lab and Pioneer Centre for AI,
University of Copenhagen, Denmark

Lewis Hammond LEWIS.HAMMOND@COOPERATIVEAI.ORG
Cooperative AI Foundation, United Kingdom

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

A key challenge for the safety of advanced AI systems is the possibility that multiple simpler agents might inadvertently form a collective agent with capabilities and goals distinct from those of any individual. More generally, determining when a group of agents can be viewed as a unified collective agent is a foundational question in the study of interactions and incentives in both biological and artificial systems. We adopt a behavioral perspective in answering this question, ascribing collective agency to a group when viewing the group's joint actions as rational and goal-directed successfully predicts its behavior. We formalize this perspective on collective agency using causal games (Hammond et al., 2023) – which are causal models of strategic, multi-agent interactions – and causal abstraction (Rubenstein et al., 2017; Beckers and Halpern, 2019) – which formalizes when a simple, high-level model faithfully captures a more complex, low-level model. We use this framework to solve a puzzle regarding multi-agent incentives in actor-critic models and to make quantitative assessments of the degree of collective agency exhibited by different voting mechanisms. Our framework aims to provide a foundation for theoretical and empirical work to understand, predict, and control emergent collective agents in multi-agent AI systems.

Keywords: Collective Agency, Causal Abstraction, Causal Incentives, AI Safety

1. Introduction

When does it make sense to view a group of individual agents as a unified collective agent? While this question may seem rather subjective at first, there is an intuitive sense in which we might ascribe collective agency to, for example, a well-organized, efficient, and synchronized team, as opposed to a random assortment of individuals pursuing independent courses of action. This is despite the fact that such a team may not have a leader or hierarchical structure (such as a jazz ensemble improvising, or a school of fish avoiding a predator), and may not even have the same objectives as one another (such as a market of buyers and sellers, or pollinating insects and flowering plants).

Indeed, humans often use notions of collective agency to understand the world (Roth, 2017; Schweikard and Schmid, 2021). For example, it is not uncommon for people to use language such as "country A wants X, which conflicts with the interests of country B" or "company C is

© 2026 F.H. Jørgensen, S. Weichwald & L. Hammond.

pursuing a new strategy Y". While statements like these are usually supposed to be approximations, they can still be helpful for understanding the world and making predictions about the future. Some have even gone so far as to argue that individual minds are composed of agentic subsystems (Nagel, 1971; Minsky, 1988), in which case there is no privileged level of abstraction on which to locate agency. How can we make sense of agency arising at various levels of abstraction?

Far from being purely philosophical, this question has important implications for understanding and predicting collective behavior in multi-agent systems, and hence for ensuring the safety of networks of artificial agents. A key safety challenge here is the possibility that multiple simpler AI systems might inadvertently form a 'super-agent' with capabilities and goals distinct from any individual in the group (Drexler, 2019; Hammond et al., 2025). Compared to AI tools, the ability of artificial *agents* to make plans and take actions in pursuit of complex goals makes them not only more useful, but also more harmful if those goals are misaligned or their pursuit produces negative side effects (Chan et al., 2023). For example, competitive pressures may lead individually rational AI agents to rapidly exhaust collective resources (Piatti et al., 2024), or a group of agents might combine their harmless individual capabilities to override safeguards and execute a cyberattack (Jones et al., 2024).

1.1. Example: Actor-Critic Agents

As a simple example of a collective agent, consider the well-known family of reinforcement learning (RL) agents known as *actor-critic* (AC) algorithms (Konda and Tsitsiklis, 2000). An AC agent comprises a critic that attempts to quantify how well the actor is performing relative to an objective, and an actor that attempts to improve its strategy according to the critic's judgment. A single-step decision problem featuring an AC algorithm is shown as a causal game (Hammond et al., 2023) in Figure 1.

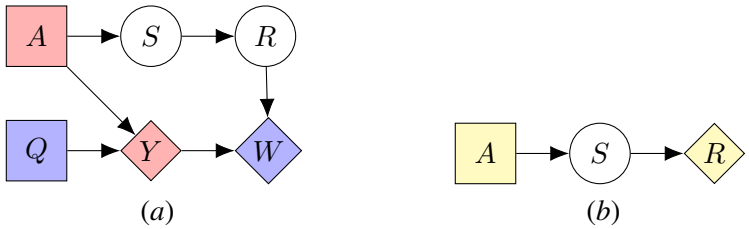

Figure 1: (a) A causal game representing an AC algorithm. In a causal game, square nodes represent decisions, circular nodes represent chance variables, diamond nodes represent utilities, and edges represent causal dependencies. In this game, the actor chooses an action a, which leads to a new state s and in turn the (true) reward r.[2] The critic, on the other hand, chooses a Q function q that takes as input the action a to produce a *predicted* reward $y = q(a)$. The actor's utility is given by y and the critic's utility is given by the loss $w = \ell(y, r)$ where ℓ is a loss function. (b) A causal influence diagram (a single-agent causal game) forming a valid abstraction of the causal game in panel (a) of Figure 1. In this game, the (collective) AC agent effectively acts so as to maximize the reward directly. Both figures are adapted from Kenton et al. (2023).

2. Note that in this simple, single-step AC problem, we leave the (fixed) starting state as implicit, such that Y only has A as a parent. While it is common for the reward function to be a function of the current state, action, and the new state,

Even though neither the actor nor the critic is maximizing for the reward, the overall system still intuitively seems to pursue this goal. In this work, we argue that this is precisely because there is a valid *abstraction* that faithfully represents the low-level causal game (Rubenstein et al., 2017; Beckers and Halpern, 2019; Beckers et al., 2020), and which corresponds to the overall agent (shown in panel (b) of Figure 1). While Kenton et al. (2023) also suggested that panel (b) of Figure 1 was a possible alternative model of the true AC dynamics in panel (a), they made no argument as to why. We solve this problem, providing a formal account of agency at different levels of abstraction.

1.2. Related Work

The question of when multiple individuals constitute a unified collective agent has long occupied philosophers, with foundational work establishing the importance of collective intentionality and shared intentions (Searle, 1990; Tuomela, 2006; Ludwig, 2007; List and Pettit, 2011; Pacherie, 2013; Bratman, 2014). While most of these prior works focus on a rich notion of collective agency, such as requiring the mental modeling of collaborators, we take a behavioral or 'black-box' approach, reducing the question of whether something is a collective agent to whether ascribing collective agency is predictive of the group's behavior. Though these works provide conceptual foundations, they generally lack the formal machinery needed to rigorously identify collective agents. Recent work has made progress on defining *individual* agents using causal and decision-theoretic tools (Orseau et al., 2018; Kenton et al., 2023; MacDermott et al., 2024; Xu and Rivera, 2024; Abel et al., 2025; Everitt et al., 2025; Rajcic and Søgaard, 2025), but none formally consider *collective* agents.

For multi-agent settings, Hammond et al. (2023) introduce causal games – a framework generalizing multi-agent influence diagrams (Koller and Milch, 2003) to higher levels of Pearl's causal hierarchy (Pearl, 2009). Other game-theoretic approaches provide models of coalition formation (Ray, 2007; Elkind and Rothe, 2016) but do not address whether such coalitions constitute unified agents. The concept of causal abstraction (Rubenstein et al., 2017; Beckers and Halpern, 2019; Beckers et al., 2020) formalizes when a high-level causal model is a valid abstraction of a detailed low-level model, preserving causal relationships when variables are aggregated. However, existing causal abstraction work focuses on causal models without agents, requiring extension to multi-agent settings with strategic interactions. A discussion of further related work can be found in Appendix A.

1.3. Contributions

After reviewing the necessary background in Section 2, we first use causal games (Hammond et al., 2023) to define what we mean by an individual decision-making agent, according to a given notion of rationality (such as playing a best response), in Section 3. Then, in Section 4 we leverage the concept of causal abstraction to provide a principled account of when a set of 'low-level' agents can be usefully modeled as a single 'high-level' collective agent. Using these formal tools, we prove a proposition that rules out the emergence of non-trivial agency at higher levels of abstraction and analyze the emergence of collective agency in the actor-critic example (Section 1.1). We further illustrate this framework in Section 5 by empirically analyzing different voting mechanisms and the extent to which they transform the different voters into a collective agent. Section 6 concludes with a brief summary and an overview of directions for future work. Appendix H further discusses

including sufficient information within the new state leads to an equivalent formulation where the reward is simply a function of the variable S.

potential applications of our framework to multi-agent reinforcement learning and large language model agents.

2. Background

Our notation and setup draw inspiration from Geiger et al. (2025). We begin with the notion of a signature, which describes a set of variables and the values they may take.

Definition 1 *A **signature** is a tuple $\Sigma = (\boldsymbol{X}, (\mathrm{ran}(X_i))_{i\in[d]})$, where $\boldsymbol{X}$ is a tuple of variables $(X_1, \ldots, X_d)$ and $(\mathrm{ran}(X_i))_{i\in[d]}$ is a tuple of ranges for each variable.*[3]

For any subset $\boldsymbol{Y} \subseteq \boldsymbol{X}$, we define $\mathrm{ran}(\boldsymbol{Y}) = \times_{Y\in\boldsymbol{Y}} \mathrm{ran}(Y)$. Formally, we assume that the variables have disjoint ranges.[4] Therefore, we may consider the elements of $\mathrm{ran}(\boldsymbol{Y})$ as sets rather than tuples. This technical assumption allows us to identify settings of variables $\boldsymbol{Y} \subseteq \boldsymbol{X}$ with sets $\boldsymbol{y} \in \mathrm{ran}(\boldsymbol{Y})$, which is useful for defining projections (Definition 2) and mechanized abstractions (Definition 14).

Definition 2 *Given two sets of variables $\boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{X}$, and setting $\boldsymbol{y} \in \mathrm{ran}(\boldsymbol{Y})$, we define the **projection** $\Pi_{\boldsymbol{Z}}(\boldsymbol{y}) = \boldsymbol{y} \cap (\bigcup_{Z\in\boldsymbol{Z}} \mathrm{ran}(Z))$.*

For example, consider a signature given by the real-valued variables $\boldsymbol{X} = \{X_1, X_2\}$, i.e., where each has range $\{x_{X_i} \mid x \in \mathbb{R}\}$ for $i \in 1, 2$. Then $\Pi_{\{X_1\}}(\{4_{X_1}, 5_{X_2}\}) = \{4_{X_1}\}$ and $\Pi_{\{X_1\}}(\{5_{X_2}\}) = \varnothing$.

Definition 3 *Given a signature Σ, a **deterministic (cyclic) structural causal model** $\widetilde{\mathcal{M}}$ over Σ is a set of functions $\{\mathcal{F}_X\}_{X\in\boldsymbol{X}}$, where $\mathcal{F}_X : \mathrm{ran}(\boldsymbol{X}\backslash\{X\}) \to \mathrm{ran}(X)$. By $\mathsf{S}(\widetilde{\mathcal{M}})$ we denote the elements $\boldsymbol{x} \in \mathrm{ran}(\boldsymbol{X})$ – i.e. the **solutions** – that satisfy $\mathcal{F}_X(\boldsymbol{x}\backslash\Pi_X(\boldsymbol{x})) = \Pi_X(\boldsymbol{x})$ for all $X \in \boldsymbol{X}$.*

Notice that while deterministic structural causal models may be cyclic, we will assume that probabilistic causal models are acyclic (see Definition 5).

Definition 4 *A **hard intervention** in a deterministic SCM is a setting $\boldsymbol{y}$ of some subset of variables $\boldsymbol{Y} \subseteq \boldsymbol{X}$. By $\mathsf{S}(\widetilde{\mathcal{M}}; \boldsymbol{y})$ – the **solutions** in $\widetilde{\mathcal{M}}$, given intervention $\boldsymbol{y}$ – we denote the set of values $\boldsymbol{x} \in \mathrm{ran}(\boldsymbol{X})$ such that $\mathcal{F}_X(\boldsymbol{x}\backslash\Pi_X(\boldsymbol{x})) = \Pi_X(\boldsymbol{x})$ for $X \notin \boldsymbol{Y}$ and $\Pi_X(\boldsymbol{x}) = \Pi_X(\boldsymbol{y})$ for $X \in \boldsymbol{Y}$.*

A hard intervention corresponds to replacing the functions $\mathcal{F}_X$ with $\boldsymbol{z} \mapsto \Pi_X(\boldsymbol{y})$ for $X \in \boldsymbol{Y}$, and the solutions $\mathsf{S}(\widetilde{\mathcal{M}}; \boldsymbol{y})$ correspond to the settings $\boldsymbol{x}$ of variables $\boldsymbol{X}$ that are solutions to this new system of equations.

Definition 5 *Given a signature Σ, a **structural causal model** (SCM) $\mathcal{M}$ over Σ is a tuple*

$$((\boldsymbol{V}, \boldsymbol{\mathcal{E}}), \mathcal{G}, \{\mathcal{F}_V\}_{V\in\boldsymbol{V}}, \mathbb{P}(\boldsymbol{\mathcal{E}}))\,,$$

where

- *$(\boldsymbol{V}, \boldsymbol{\mathcal{E}})$ is a partition of the variables in Σ into noise variables $\boldsymbol{\mathcal{E}}$ and object variables $\boldsymbol{V}$. We assume that each object variable has a unique noise variable associated with it, that is, $\boldsymbol{\mathcal{E}} = \{\mathcal{E}_V\}_{V\in\boldsymbol{V}}$.*

3. We sometimes consider $\boldsymbol{X}$ a set rather than a tuple.
4. The assumption of disjoint ranges can be made without loss of generality: we subscript every value with the variable name to ensure uniqueness, such that, for example, $x_{X_i} \neq x_{X_j}$ for $x \in \mathbb{R}$ and $i \neq j$. If it is clear from the context which range a value belongs to, we may omit the subscript.

- *$\mathcal{G}$ is a directed acyclic graph over the object variables $\boldsymbol{V}$.*
- *$\{\mathcal{F}_V\}_{V\in\boldsymbol{V}}$ is a set of structural assignments $\mathcal{F}_V : \mathrm{ran}(\mathbf{PA}_V^{\mathcal{G}} \cup \{\mathcal{E}_V\}) \to \mathrm{ran}(V)$, where $\mathbf{PA}_V^{\mathcal{G}}$ denotes the parents of V in $\mathcal{G}$.*
- *$\mathbb{P}(\boldsymbol{\mathcal{E}})$ is a distribution on $\mathrm{ran}(\boldsymbol{\mathcal{E}})$. We assume that the variables $\{\mathcal{E}_V\}_{V\in\boldsymbol{V}}$ are jointly independent.*

An SCM $\mathcal{M}$ induces a unique distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{V})$ over the object variables $\boldsymbol{V}$ such that $V = \mathcal{F}_V(\mathbf{PA}_V, \mathcal{E}_V)$, in distribution, for every $V \in \boldsymbol{V}$. Since the noise variables are jointly independent, the induced distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{V})$ is Markovian with respect to $\mathcal{G}$ (Peters et al., 2017, Prop. 6.31).

In order to consider causal games in which the mechanisms (i.e. the structural functions and noise variables) governing an object variable can change if an agent selects a different strategy, it will be useful to define parameterized SCMs, which is a class of SCMs with functions that are indexed by parameters.[5]

Definition 6 *Given a signature Σ, a **parameterized SCM** $\mathcal{M}^{\boldsymbol{\Theta}}$ over Σ is a tuple*

$$\Big((\boldsymbol{V}, \boldsymbol{\mathcal{E}}), \mathcal{G}, \{\Theta_V\}_{V\in\boldsymbol{V}}, \{\mathcal{F}_V^{\theta}\}_{V\in\boldsymbol{V},\theta\in\Theta_V}, \mathbb{P}(\boldsymbol{\mathcal{E}})\Big)$$

where $(\boldsymbol{V}, \boldsymbol{\mathcal{E}}), \mathcal{G}$, and $\mathbb{P}(\boldsymbol{\mathcal{E}})$ are as for an SCM (Definition 5), and

- *$\{\Theta_V\}_{V\in\boldsymbol{V}}$ is a set of parameter spaces, one for each object variable.*
- *$\{\mathcal{F}_V^{\theta}\}_{V\in\boldsymbol{V},\theta\in\Theta_V}$ is a set of structural assignments $\mathcal{F}_V^{\theta} : \mathrm{ran}(\mathbf{PA}_V^{\mathcal{G}} \cup \{\mathcal{E}_V\}) \to \mathrm{ran}(V)$, which are indexed by variables V and parameters $\theta_V \in \Theta_V$.*

For a given setting $\boldsymbol{\theta} = \{\theta_V\}_{V\in\boldsymbol{V}}$ of the parameters $\theta_V \in \Theta_V$, a parameterized SCM induces an SCM $\mathcal{M}^{\boldsymbol{\theta}}$.

We are now ready to provide a formal definition of *mechanized* structural causal models (Hammond et al., 2023), in which the parameters θ_V of each structural function (representing the causal mechanisms in the model) are explicitly represented and governed by a separate deterministic cyclic SCM.[6] We depart slightly from Hammond et al. (2023) by having the mechanisms being governed by functions rather than relations.

Definition 7 *A **mechanized SCM** $m\mathcal{M}$ is a tuple $(\widetilde{\mathcal{M}}, \mathcal{M}^{\boldsymbol{\Theta}})$ where*

- *$\widetilde{\mathcal{M}}$ is a deterministic cyclic SCM with mechanism variables $\widetilde{\boldsymbol{V}}$.*
- *$\mathcal{M}^{\boldsymbol{\Theta}}$ is a parameterized SCM with object variables $\boldsymbol{V}$.*

*We refer to $\boldsymbol{V}$ as the **object nodes** and to $\widetilde{\boldsymbol{V}}$ as the **mechanism nodes**. We assume that there is a one to one correspondence between the mechanism variables $\widetilde{\boldsymbol{V}}$ and object variables $\boldsymbol{V}$, and that $\mathcal{M}^{\boldsymbol{\Theta}}$'s parameter sets are $\Theta_V = \mathrm{ran}(\widetilde{V})$ for each $V \in \boldsymbol{V}$. Every solution $\boldsymbol{s} \in \mathcal{S}(\widetilde{\mathcal{M}})$ induces an SCM $\mathcal{M}^{\boldsymbol{s}}$ with structural assignments $\{\mathcal{F}_V^{\Pi_{\widetilde{V}}(\boldsymbol{s})} \mid V \in \boldsymbol{V}\}$ and distribution $\mathbb{P}_{\boldsymbol{s}}(\boldsymbol{V})$. By $\mathbb{P}_{\mathcal{S}(\widetilde{\mathcal{M}})}(\boldsymbol{V})$ we denote the set of distributions $\{\mathbb{P}_{\boldsymbol{s}}(\boldsymbol{V}) \mid \boldsymbol{s} \in \mathcal{S}(\widetilde{\mathcal{M}})\}$.*

5. This restriction to parameterized functions – again departing from (Hammond et al., 2023) – is more natural for AI agents whose strategies are represented by parametric models such as a neural network.
6. Dawid (2002) provides an earlier example of mechanized SCMs, though he does not consider the mechanism nodes themselves as being governed by an SCM, rather each mechanism node is parentless.

To help the reader understand this formalism, we provide a worked out formalization of the 'Battle of the Sexes' game as a mechanized SCM in Appendix B.

3. Agency

We define utilities as functions of nodes (i.e. object variables) rather than nodes in the causal graph itself, highlighting the constitutive rather than causal relationship between utilities and object variables (Xia and Bareinboim, 2024). For example, we may have two different agents, who both care about the number of bicycles in the world, but one agent wants to maximize the number of bicycles in the world, while the other agent wants to minimize the number of bicycles in the world. In that case, it would be problematic to have two different utility nodes because there would be a logical rather than causal relationship between the two utility nodes, which would violate the assumption of independent causal mechanisms (Peters et al., 2017; Jørgensen et al., 2025).

Definition 8 *Given a mechanized SCM $m\mathcal{M}$ with object variables $\boldsymbol{V}$, we define a **utility function** as a function $\mathcal{U} : \mathrm{ran}(\boldsymbol{V}) \to \mathbb{R}$.*

While we technically allow that a utility function $\mathcal{U}$ may depend on several object variables $\mathbf{U} \subseteq \mathbf{V}$, we will usually consider utilities that depend on a single object variable U. Given a utility function, we may wonder if any object level variables can be viewed as optimizing it. For this purpose, we use rationality relations (Hammond et al., 2023), which describe how an agent would react in a given context if employing a certain conception of rationality and having a certain utility function.

Definition 9 *Given a mechanized SCM $m\mathcal{M}$, a mechanism node $\widetilde{S} \in \widetilde{\boldsymbol{V}}$ and a utility function $\mathcal{U} : \mathrm{ran}(\boldsymbol{V}) \to \mathbb{R}$, a **rationality relation** is a total relation $\mathcal{R}_{\widetilde{S}} \subseteq \mathrm{ran}(\widetilde{\boldsymbol{V}} \backslash \{\widetilde{S}\}) \times \mathrm{ran}(\widetilde{S})$. If $\mathcal{F}_{\widetilde{S}}(\widetilde{\boldsymbol{c}}) \in \mathcal{R}_{\widetilde{S}}(\widetilde{\boldsymbol{c}})$ [7] for every context $\widetilde{\boldsymbol{c}} \in \mathrm{ran}(\widetilde{\boldsymbol{V}} \backslash \{\widetilde{S}\})$, then we say that $\widetilde{S}$ responds $\mathcal{R}$-rationally to utility $\mathcal{U}$.*

Intuitively, $\mathcal{R}_{\widetilde{S}}(\widetilde{\boldsymbol{c}})$ specifies the set of policies an agent with decision node S employing rationality relation $\mathcal{R}_{\widetilde{S}}$ could employ in context $\widetilde{\boldsymbol{c}}$ if having utility function $\mathcal{U}$.

Definition 10 *Given a mechanism node $\widetilde{S} \in \widetilde{\boldsymbol{V}}$ and a utility function $\mathcal{U} : \mathrm{ran}(\boldsymbol{V}) \to \mathbb{R}$, we define the **best response rationality relation** $\mathcal{R}^{BR}_{\widetilde{S}}$ under utility $\mathcal{U}$ by the condition that*

$$\widetilde{s} \in \mathcal{R}^{BR}_{\widetilde{S}}(\widetilde{\boldsymbol{c}}) \quad \textit{if and only if} \quad \widetilde{s} \in \underset{\widetilde{s}' \in \mathrm{ran}(\widetilde{S})}{\arg\max}\, \mathbb{E}_{\widetilde{s}', \widetilde{\boldsymbol{c}}}(\mathcal{U}(\boldsymbol{V}))$$

for every $\widetilde{\boldsymbol{c}} \in \mathrm{ran}(\widetilde{\boldsymbol{V}} \backslash \{\widetilde{S}\})$.

Best response rationality is the main rationality concept we consider. Sometimes, other rationality concepts, such as subgame perfection, may be more appropriate (Hammond et al., 2023). Different notions of rationality may be particularly relevant in the context of collective agency for AI agents, where agents may be interacting with copies of themselves or have access to each other's parameters (Critch et al., 2022; Oesterheld et al., 2023). We leave a detailed investigation of various rationality relations for future work, but in Appendix C, we show that sometimes we do not have collective agency under best response rationality, while we do have it for other rationality concepts.

7. By $\mathcal{R}_{\widetilde{S}}(\widetilde{\boldsymbol{c}})$ we denote the set $\{\widetilde{s} \in \mathrm{ran}(\widetilde{S}) \mid (\widetilde{\boldsymbol{c}}, \widetilde{s}) \in \mathcal{R}_{\widetilde{S}}\}$.

Below we operationalize agency as rational, goal-driven behavior. While we do not claim that this is all that there is to the philosophical concept of agency, we do believe that this provides a useful starting point for a formalization of agency that works across levels of abstraction.

Definition 11 *Given a mechanized SCM $m\mathcal{M}$, we say a mechanism variable $\widetilde{S} \in \widetilde{\boldsymbol{V}}$ is an $(\mathcal{R},\mathcal{U})$-**agent** if $\widetilde{S}$ responds $\mathcal{R}$-rationally to utility $\mathcal{U}$.*

Kenton et al. (2023) discuss assumptions that make it possible to identify which nodes correspond to best-response rational agents and which nodes enter into their utility functions. Notice that every object variable is an agent with respect to best response rationality and a constant utility function. This is related to a classical insight in inverse reinforcement learning (Ng et al., 2000; Skalse and Abate, 2024), namely that reward functions are not generally identifiable from behavior. Likewise, if we allow any utility function, then any mechanism variable can be viewed as an agent. While modeling something as an agent optimizing for a constant utility function does not make false predictions, it also does not make any non-vacuous predictions.

Definition 12 *We say that a mechanism variable $\widetilde{S} \in \widetilde{\boldsymbol{V}}$ is a **non-trivial** $(\mathcal{R},\mathcal{U})$**-agent** if it is an $(\mathcal{R},\mathcal{U})$-agent and there exist two settings $\tilde{\boldsymbol{c}},\tilde{\boldsymbol{c}}' \in \mathrm{ran}(\widetilde{\boldsymbol{V}}\backslash\{\widetilde{S}\})$ such that $\mathbb{P}_{\tilde{\boldsymbol{c}},\mathcal{F}_{\widetilde{S}}(\tilde{\boldsymbol{c}})}(S \mid \mathbf{PA}_S) \neq \mathbb{P}_{\tilde{\boldsymbol{c}}',\mathcal{F}_{\widetilde{S}}(\tilde{\boldsymbol{c}}')}(S \mid \mathbf{PA}_S)$*

4. Collective Agency

4.1. Mechanized Abstractions

In this section, we refer to objects related to the high-level model using superscripts *. Given a low-level mechanized SCM m$\mathcal{M}$, and a corresponding high-level mechanized SCM m$\mathcal{M}^*$, we need to specify which of the low-level object variables $\boldsymbol{V}$ correspond to which high-level object variables V^*. We call this correspondence a variable alignment.

Definition 13 *Given two mechanized causal graphs $m\mathcal{M}$ and $m\mathcal{M}^*$, a **variable alignment** is a mapping $\mathtt{A} : \boldsymbol{V}^* \to \mathcal{P}(\boldsymbol{V})\backslash\{\varnothing\}$ such that $\mathtt{A}_{V_1^*} \cap \mathtt{A}_{V_2^*} = \varnothing$ for any two variables $V_1^*, V_2^* \in \boldsymbol{V}^*$ (where $\mathtt{A}_{V_1^*} := \mathtt{A}(V_1^*)$). Given a variable alignment, we define $\mathtt{A}_{\widetilde{V}^*} \subseteq \widetilde{\boldsymbol{V}}$ as those mechanism nodes that correspond to the object variables in $\mathtt{A}_{V^*}$.*

Given a variable alignment $\mathtt{A}$, a set of functions $\{\tau_{V^*}\}_{V^* \in \boldsymbol{V}^*}$, where $\tau_{V^*} : \mathrm{ran}(\mathtt{A}_{V^*}) \to \mathrm{ran}(V^*)$, induces a function $\tau : \mathrm{ran}(\boldsymbol{V}) \to \mathrm{ran}(\boldsymbol{V}^*)$ given by $\tau(\boldsymbol{v}) = \bigcup_{V^* \in \boldsymbol{V}^*} \tau_{V^*}(\Pi_{\mathtt{A}_{V^*}}(\boldsymbol{v}))$. We call these functions **value mappings**. These functions determine how we translate values of low-level object variables to values of high-level object variables. In addition, we need to specify how we translate interventions on the low-level mechanisms variables to interventions on the high-level mechanisms variables. These **intervention mappings** are given by partial functions $\{\omega_{\widetilde{V}^*}\}_{\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*}$, where $\omega_{\widetilde{V}^*} : \mathrm{ran}(\mathtt{A}_{\widetilde{V}^*}) \rightharpoonup \mathrm{ran}(\widetilde{V}^*)$. By $\mathrm{dom}(\omega_{\widetilde{V}^*})$, we denote the subset of $\mathrm{ran}(\mathtt{A}_{\widetilde{V}^*})$ on which $\omega_{\widetilde{V}^*}$ is defined. For a subset $\widetilde{\boldsymbol{Y}}^* \subseteq \widetilde{\boldsymbol{V}}^*$, we define $\mathrm{dom}(\omega_{\widetilde{\boldsymbol{Y}}^*}) = \times_{\widetilde{V}^* \in \widetilde{\boldsymbol{Y}}^*} \mathrm{dom}(\omega_{\widetilde{V}^*})$. The set of intervention mappings naturally induces a partial function ω, from any setting of a subset of low-level collections $\bigcup_{\widetilde{V}^* \in \widetilde{\boldsymbol{Y}}^*} \mathtt{A}_{\widetilde{V}^*}$ to a setting of $\widetilde{\boldsymbol{Y}}^*$, in particular, given $\tilde{\boldsymbol{y}} \in \mathrm{dom}(\omega_{\widetilde{\boldsymbol{Y}}^*})$, ω maps to $\bigcup_{\widetilde{V}^* \in \widetilde{\boldsymbol{Y}}^*} \omega_{\widetilde{V}^*}(\Pi_{\mathtt{A}_{\widetilde{V}^*}}(\tilde{\boldsymbol{y}}))$.

Definition 14 *Assume that we have two mechanized causal models $m\mathcal{M}$ and $m\mathcal{M}^*$. Let a node alignment $\mathtt{A}$ be given. We say that $m\mathcal{M}^*$ is a **mechanized abstraction** of $m\mathcal{M}$ under value mappings*

$\{\tau_{V^*}\}_{V^* \in \boldsymbol{V}^*}$ ($\tau_{V^*} : \text{ran}(A_{V^*}) \to \text{ran}(V^*)$), *and intervention mappings* $\{\omega_{\widetilde{V}^*}\}_{\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*}$ ($\omega_{\widetilde{V}^*} : \text{ran}(A_{\widetilde{V}^*}) \rightharpoonup \text{ran}(\widetilde{V}^*)$), *if*

$$\mathbb{P}_{\mathcal{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}})}(\tau(\boldsymbol{V})) = \mathbb{P}_{\mathcal{S}(\widetilde{\mathcal{M}}^*;\omega(\widetilde{\boldsymbol{y}}))}(\boldsymbol{V}^*), \tag{1}$$

for any subset $\widetilde{\boldsymbol{Y}}^* \subseteq \widetilde{\boldsymbol{V}}^*$, *and intervention* $\widetilde{\boldsymbol{y}} \in dom(\omega_{\widetilde{\boldsymbol{Y}}^*})$.

Notice that Definition 14 in particular requires observational consistency, that is, $\mathbb{P}_{\mathcal{S}(\widetilde{\mathcal{M}})}(\tau(\boldsymbol{V})) = \mathbb{P}_{\mathcal{S}(\widetilde{\mathcal{M}}^*)}(\boldsymbol{V}^*)$, by considering the empty subset $\widetilde{\boldsymbol{Y}}^* = \varnothing$.

Analogously to Beckers and Halpern (2019), we define strong mechanized abstractions by adding the requirement that for any setting of a subset of high-level mechanisms, there exists a low-level intervention that implements it.

Definition 15 *We say that* $m\mathcal{M}^*$ *is a* ***strong mechanized abstraction*** *of* $m\mathcal{M}$ *under* $\{\tau_{V^*}\}_{V^* \in \boldsymbol{V}^*}$ *and* $\{\omega_{\widetilde{V}^*}\}_{\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*}$ *if* $m\mathcal{M}^*$ *is a mechanized abstraction of* $m\mathcal{M}$ *under* $\{\tau_{V^*}\}_{V^* \in \boldsymbol{V}^*}$ *and* $\{\omega_{\widetilde{V}^*}\}_{\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*}$, *and each* $\omega_{\widetilde{V}^*}$ *is surjective onto* $\text{ran}(\widetilde{V}^*)$.

If the mechanism models $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{M}}^*$ have no edges, then the definition above is an extension of constructive abstractions (Beckers and Halpern, 2019) that includes soft interventions a (hard intervention on mechanism variables may correspond to a soft intervention on the object variables). Unlike previous work on abstractions of soft interventions (Massidda et al., 2023), we do not impose any restrictions on ω except that it is induced from intervention mappings on the variable alignment. We believe that any further restrictions on ω must derive from context-specific considerations rather than being part of the definition. There are several ways in which this definition can be extended to approximate abstractions (Beckers et al., 2020; Rischel and Weichwald, 2021) and we likewise believe that the appropriate generalization is context-dependent. In Section 5, we consider restrictions on ω and a notion of approximate abstraction based on mean squared error for a fixed distribution over low-level interventions. In other applications, it is likely appropriate to consider KL divergence as is done, for example, in Dyer et al. (2024).

4.2. Ruling out emergence of non-trivial agency

We may want to rule out the emergence of agency at higher levels of abstraction. In this section, we provide a first example of a theorem that rules out that a node is a high-level agent based on properties of the low level.

We say that a mechanism node $\widetilde{S}$ has an **independent mechanism** if there exists a setting $\widetilde{s} \in \text{ran}(\widetilde{S})$ such that $\mathcal{F}_{\widetilde{S}}(\widetilde{\boldsymbol{c}}) = \widetilde{s}$ for all $\widetilde{\boldsymbol{c}} \in \text{ran}(\widetilde{\boldsymbol{V}} \backslash \{\widetilde{S}\})$.

Proposition 16 *Let mechanized models* $m\mathcal{M}$ *and* $m\mathcal{M}^*$ *be given. Assume that* $m\mathcal{M}^*$ *is a strong mechanized abstraction of* $m\mathcal{M}$ *for some value mappings* $\{\tau_{V^*}\}_{V^* \in \boldsymbol{V}^*}$ *and intervention mappings* $\{\omega_{\widetilde{V}^*}\}_{\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*}$. *Let* $\widetilde{S}^* \in \widetilde{\boldsymbol{V}}^*$ *be a high-level mechanism variable. Assume that (i)* $\tau_{\mathbf{PA}_{S^*}}$ *is injective, and (ii) the mechanism nodes in* $A_{\widetilde{S}^*}$ *have independent mechanisms. Then* $\widetilde{S}^*$ *is not a non-trivial agent in* $m\mathcal{M}^*$. *[See Appendix E for the proof.]*

We think that having a more complete theory of the (non-)emergence of agency is an important direction for future work.

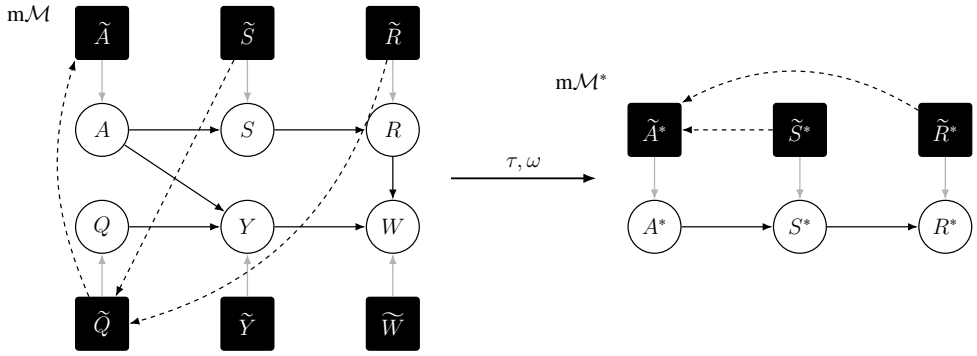

Figure 2: A graphical representation of a causal abstraction applied to mechanized SCMs representing the AC example from Section 1.1 (see Appendix D for the explicit construction of the abstraction). On the left hand side, we have the low-level model m$\mathcal{M}$. On the right, we have a high-level model m$\mathcal{M}^*$, which is a valid abstraction of the low-level model under τ and ω. By examining the high-level model, we can see that A^* is an $\mathcal{R}^{\text{BR}}$-rational agent with utility equal to the reward: $\mathcal{U}(\boldsymbol{v}^*) = \Pi_{R^*}(\boldsymbol{v}^*)$.

4.3. Example: Actor-Critic Agents (Revisited)

We revisit the one-step AC agent (as described in Kenton et al. (2023) and Section 1.1). Here the critic $\widetilde{Q}$ is tasked with predicting the reward R for different actions that the actor $\widetilde{A}$ may take. The action A taken by the actor $\widetilde{A}$ affects the state, and in turn, the reward. Kenton et al. (2023) hypothesize that one can, at a more coarse-grained level, view this as a single-agent system. This in turn explains why the system intuitively has an incentive to manipulate the state S, even though if one applies the graphical criterion for single-agent graphs (Everitt et al., 2021), it suggests that neither $\widetilde{A}$ nor $\widetilde{Q}$ have an incentive to manipulate the state S. See Appendix D for details about how to formalize this setup in terms of mechanized SCMs.

Proposition 17 *In Figure 2 (see also Appendix D for details), m$\mathcal{M}^*$ is a strong mechanized abstraction of m$\mathcal{M}$. Furthermore, $\widetilde{A}^*$ in the high-level model m$\mathcal{M}^*$ is an $(\mathcal{R}^{BR}, \mathcal{U})$-agent with utility equal to the reward $\mathcal{U}(\boldsymbol{v}^*) = \Pi_{R^*}(\boldsymbol{v}^*)$. [See Appendix F for the proof.]*

Note that while the AC example 'marginalizes' out one of the agents (namely the critic), our framework allows for more general types of collective agency. For example, in the next section, we consider aggregating individuals into a single agent.

5. Example: Using Collective Agency for Surrogate Modeling

In this section we show an application of collective agency for surrogate modeling. Specifically, we investigate the degree to which it makes sense to model countries as collective agents under various voting mechanisms. Dyer et al. (2024) introduced the idea of using causal abstractions for interventionally consistent surrogate modeling. We build on this work by extending it to mechanized abstractions with agents. There are at least three potential benefits of using surrogate models with collective agents: 1) the low-level model is often a black box where new interventions cannot be

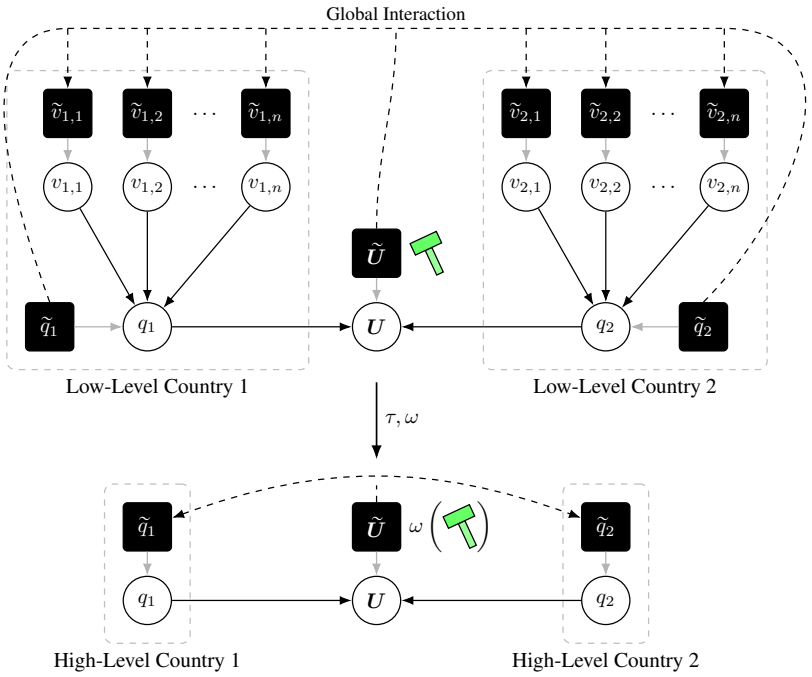

Figure 3: At the low level, the global interaction includes edges between every pair of citizens and from the voting mechanism $\widetilde{q}_1$ and $\widetilde{q}_2$ to every citizen. At the high level, we view the countries as being individual agents, picking the pollution level in response to the other countries. Since we are intervening on the citizens preferences, we explicitly include a (vector valued) utility node $\widetilde{\boldsymbol{U}}$. At the low level, it has dimension equal to the number of citizens. At the high level, it has dimension equal to the number of countries. In the figure and the main text, we omit superscripts (*) for the high level.

simulated;[8] 2) computing low-level equilibria may be computationally expensive, which hinders evaluations of new interventions; and 3) agency-based surrogate models may be useful for assessing abstract counterfactuals such as "how would country A react if country B adopts policy X?".

While we believe that there are practical applications where surrogate modeling with agents is useful, the main point of the example below is to provide a theoretical illustration and proof of concept. We therefore prioritize simplicity over realistic assumptions.

5.1. The Low-Level Model

We have C countries and each country c has N_c citizens. Each country employs a voting mechanism to determine the level of pollution q_c. We assume that individual i from country c has utility

$$U_{ci}(q_c, q_{-c}) = a_{ci}q_c - b_{ci}q_c^2 - d_{ci}Q_W^2,$$

8. This is essentially the point of discovering psychological regularities. It is much easier to predict people's behavior based on abstract models of beliefs and preferences than based on neurophysiology (Dennett, 1989).

where q_{-c} are the pollution levels of countries other than c, $a_{ci}, b_{ci}, d_{ci} \in \mathbb{R}$ are individual parameters, and $Q_W = \sum_c q_c$. We consider interventions on individuals' preferences for pollution, for example, implemented by taxation or advocacy, such that a_{ci} is replaced with $a_{ci} - \lambda_{ci} \geqslant 0$.

We assume that the result of the countries' votes is a Nash equilibrium (NE) where every citizen of every country is a player. This NE may be computationally expensive to evaluate or, if the low-level model is not understood, can only be evaluated by playing it out in the physical world. Rather than reasoning about NEs containing every citizen of every country, it may be possible to model the situations at a higher level of abstraction where countries are modeled as agents.

5.2. The High-Level Model

We use a parametric model for country utilities, which mimics the utilities of individual agents:

$$U_c(q_c, q_{-c}) = A_c q_c - B_c q_c^2 - D_c Q_W^2. \tag{2}$$

Assuming that each country is an agent with this utility function, it is easy to derive that under NE, the total amount of pollution is

$$Q_W^{\text{NE}} = \frac{\frac{1}{2}\sum_c \frac{A_c}{B_c}}{1 + \sum_c \frac{D_c}{B_c}},$$

and the individual contributions are

$$q_c^{\text{NE}} = \frac{A_c}{2B_c} - \frac{D_c}{B_c} Q_W^{\text{NE}}. \tag{3}$$

Notice that there is no guarantee that the voting mechanisms will produce pollution levels that are correctly modeled by the countries playing an NE according to a utility function parameterized by Equation (2). This is a modeling choice that may or may not be appropriate depending on the actual voting mechanism that is used and the utility functions of the individual citizens.

5.3. Connecting the Low-Level and High-Level Models

We want to learn a mapping ω from the vector intervention $\boldsymbol{\lambda} := (\lambda_{ci})_{ci}$ to an intervention on $(A_c, B_c, D_c)_c$ such that the high-level model is a mechanized abstraction of the low-level model (see Figure 3). In other words, we want to see if we can view the countries as collective agents. Formally, ω ought to be such that

$$\mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\boldsymbol{\lambda})}((q_c)_c) = \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}}^*;\omega(\boldsymbol{\lambda}))}((q_c)_c)$$

where we assume that solutions are given by best response rationality. In words, we want that applying intervention $\boldsymbol{\lambda}$ to the citizen preferences results in a low-level NE that equals the high-level NE under the $\omega(\boldsymbol{\lambda})$ intervention in the high-level model.

Notice that the high-level NE only depends on the parameters A_c,B_c, and D_c through the fractions $\alpha_c := \frac{A_c}{B_c}$ and $\delta_c := \frac{D_c}{B_c}$. A single sample is not enough to identify α_c and δ_c (for each country, we have a single equation – Equation (3) – and two unknowns). Therefore we need to make assumptions about how our intervention affects the high-level utility function. For simplicity, we assume that the interventions only affect the α_c's and not the δ_c's (which we know to be true for VCG voting; see

Appendix G.1). Whether this assumption is reasonable for other voting mechanisms is an empirical question, which we explore in the results section.

Using this assumption, we estimate ω using a two-step procedure. First, for each country c, we apply interventions such that $\lambda_{ci} = 0$ for all citizens i of country c, but non-zero for other countries. We then estimate δ_c by linear regression of q_c on Q_W (which is justified by Equation (3)). Second, we parametrize $\omega : \mathbb{R}^{\sum_c N_c} \to \mathbb{R}^C$ by the weights ϕ of a neural network and train the network to minimize the loss function

$$\text{Loss}(\phi) = \sum_j \sum_c \left[\widehat{q}_c^{\text{NE}}(\phi(\boldsymbol{\lambda}^j)) - q_c^{\text{NE}}(\boldsymbol{\lambda}^j)\right]^2,$$

where $\boldsymbol{\lambda}^j$ is an intervention on the citizens preferences, $q_c^{\text{NE}}(\boldsymbol{\lambda}^j)$ is the resulting ground-truth low-level NE, and $\widehat{q}_c^{\text{NE}}(\phi(\boldsymbol{\lambda}^j))$ is the NE derived by Equation (3) and the estimated $\widehat{\delta_c}$ and $\widehat{\alpha_c}$, the latter of which is one of the outputs returned by ω. The interventions $\boldsymbol{\lambda}^j$ are sampled according to a fixed distribution (see Appendix G.2 for details). Minimizing the consistency loss is the central idea of approximate causal abstractions (Beckers et al., 2020; Rischel and Weichwald, 2021).

5.4. Results

We consider three different voting mechanisms: Vickrey-Clarke-Groves (VCG), Median Voting, and Random Dictator. Descriptions of the three voting mechanisms and how we compute low-level Nash equilibria for each are given in Appendix G.1. We apply our method to a simulation with five countries and a total of 1,000 citizens. See Appendix G.2 for details about the simulations and training. We report a summary of the results in Table 1 and provide detailed results in Appendix G.3.

Mechanism	Model MAE	Baseline MAE	Improvement
VCG	0.062	1.254	95.0%
Median	0.142	1.248	88.6%
Random Dictator	5.077	4.645	-9.3%

Table 1: We are predicting q_c for different voting mechanisms under interventions $\boldsymbol{\lambda}$. (MAE) is mean absolute error for different voting mechanisms. The baseline is calculated by constantly predicting the low-level Nash equilibrium at $\lambda = 0$ (since random dictator is stochastic we use the average low-level NE for this mechanism as the baseline). The errors are summed across all countries. VCG and Median voting show low errors (compared to the baseline), indicating that the collective agency framework effectively captures how interventions affect outcomes. Random Dictator performs worse than the baseline (-9.3%) and has large absolute errors, demonstrating the difficulty of learning an ω for this mechanism.

6. Summary and future work

In this work we have laid out a mathematical foundation of agency and collective agency based on the mathematical formalisms of mechanized causal graphs and causal abstraction. The framework opens up many interesting questions that we hope can be addressed in future work, including: 1) Investigating how different notions of rationality, such as different decision theories, affect the propensity for collective agency; 2) Developing a theory that predicts when collective agency

will emerge (see Proposition 16); 3) Extending inverse game theory or inverse reinforcement learning to various levels of abstraction; and 4) Translating our theory of collective agency into a mathematical framework that more naturally accommodates both non-independent mechanisms and causal abstractions, such as Garrabrant et al. (2024). We also hope that our work can help to provide the foundation for further theoretical and empirical work to understand, predict, and control emergent collective agents in networks of AI systems.

7. Acknowledgments

This work was initiated during FHJ's participation in the Pivotal Research Fellowship. We thank Casper Lützhøft Christensen, Soroush Ebadian, Sukanya Krishna, Francisco Madaleno, Kyle Reynoso, Morgan Simpson, and Riya Tyagi for helpful discussions and feedback.

References

David Abel, André Barreto, Michael Bowling, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, et al. Agency is frame-dependent. *arXiv*, 2025. arXiv:2502.04403. (Cited on page 3.)

P. W. Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, 1972. (Cited on page 18.)

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *AAAI Conference on Artificial Intelligence*, 2019. (Cited on pages 1, 3, 8, and 19.)

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, 2020. (Cited on pages 3, 8, 12, and 19.)

Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, 1999. (Cited on page 19.)

Michael E. Bratman. *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press, 2014. (Cited on pages 3 and 18.)

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Conference on Fairness, Accountability, and Transparency*, 2023. (Cited on pages 2 and 20.)

Chih-Chun Chen, Sylvia B. Nagl, and Christopher D. Clack. A formalism for multi-level emergent behaviours in designed component-based systems and agent-based simulations. In Moulay Aziz-Alaoui and Cyrille Bertelle, editors, *From System Complexity to Emergent Properties*, chapter 4, pages 101–114. Springer, 2009. (Cited on page 20.)

Iain D. Couzin. Collective cognition in animal groups. *Trends in Cognitive Sciences*, 13(1):36–43, 2009. (Cited on page 19.)

Andrew Critch, Michael Dennis, and Stuart Russell. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *arXiv*, 2022. arXiv:2208.07006. (Cited on page 6.)

A Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 2002. (Cited on page 5.)

Daniel C Dennett. *The intentional stance*. MIT Press, 1989. (Cited on page 10.)

K. Eric Drexler. Reframing superintelligence: Comprehensive AI services as general intelligence. Technical Report 2019-1, Future of Humanity Institute, University of Oxford, 2019. (Cited on pages 2 and 20.)

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Learning Representations*, 2024. (Cited on page 30.)

Joel Dyer, Nicholas George Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Ani Calinescu, Theodoros Damoulas, and Michael J. Wooldridge. Interventionally consistent surrogates for complex simulation models. In *Advances in Neural Information Processing Systems*, 2024. (Cited on pages 8 and 9.)

Edith Elkind and Jörg Rothe. *Cooperative Game Theory*, chapter 3, pages 135–193. Springer, 2016. (Cited on pages 3 and 19.)

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *AAAI Conference on Artificial Intelligence*, 2021. (Cited on page 9.)

Tom Everitt, Cristina Garbacea, Alexis Bellot, Jonathan Richens, Henry Papadatos, Siméon Campos, and Rohin Shah. Evaluating the goal-directedness of large language models. *arXiv*, 2025. arXiv:2504.11844. (Cited on pages 3, 19, and 31.)

Karl J Friston, Maxwell J D Ramstead, Alex B Kiefer, Alexander Tschantz, Christopher L Buckley, Mahault Albarracin, Riddhi J Pitliya, Conor Heins, Brennan Klein, Beren Millidge, Dalton A R Sakthivadivel, Toby St Clere Smithe, Magnus Koudahl, Safae Essafi Tremblay, Capm Petersen, Kaiser Fung, Jason G Fox, Steven Swanson, Dan Mapes, and Gabriel René. Designing ecosystems of intelligence from first principles. *Collective Intelligence, 3(1), 2024*, 3(1), 2022. (Cited on page 19.)

Scott Garrabrant, Matthias Georg Mayer, Magdalena Wache, Leon Lang, Sam Eisenstat, and Holger Dell. Factored space models: Towards causality between levels of abstraction. *arXiv*, 2024. arXiv:2412.02579. (Cited on page 13.)

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26, 2025. (Cited on page 4.)

Natalie Gold and Robert Sugden. Collective intentions and team agency. *Journal of Philosophy*, 104 (3):109–137, 2007. (Cited on page 18.)

Joseph Y. Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *AAAI Conference on Artificial Intelligence*, 2018. (Cited on page 19.)

Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. Reasoning about causality in games. *Artificial Intelligence*, 320:103919, 2023. (Cited on pages 1, 2, 3, 5, 6, and 19.)

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced AI. Technical Report 1, Cooperative AI Foundation, 2025. (Cited on pages 2 and 20.)

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. (Cited on page 30.)

Erik Jones, Anca Dragan, and Jacob Steinhardt. Adversaries can misuse combinations of safe models. *arXiv*, 2024. arXiv:2406.14595. (Cited on pages 2 and 20.)

Frederik Hytting Jørgensen, Luigi Gresele, and Sebastian Weichwald. What is causal about causal models and representations? *arXiv*, 2025. arXiv:2501.19335. (Cited on pages 6 and 30.)

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *Artificial Intelligence*, 2023. (Cited on pages 2, 3, 7, 9, 19, and 24.)

Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior*, 45(1):181–221, 2003. (Cited on pages 3 and 19.)

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, pages 1008–1014. MIT Press, 2000. (Cited on page 2.)

Aleš Kubík. Toward a formalization of emergence. *Artificial Life*, 9(1):41–65, 2003. (Cited on page 19.)

Hector J. Levesque, Philip R. Cohen, and José H. T. Nunes. On acting together. In *Proceedings of the 8th National Conference on Artificial Intelligence. Boston, Massachusetts, USA, July 29 - August 3, 1990, 2 Volumes*, 1990. (Cited on page 19.)

Qian Li et al. Towards scalable oversight with collaborative multi-agent debate in error detection. *arXiv preprint*, 2025. (Cited on page 30.)

Christian List and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011. (Cited on pages 3 and 18.)

Kirk Ludwig. Collective intentional behavior from the standpoint of semantics. *Noûs*, 41(3):355–393, 2007. (Cited on pages 3 and 18.)

Matt MacDermott, James Fox, Francesco Belardinelli, and Tom Everitt. Measuring goal-directedness. In *Advances in Neural Information Processing Systems*, 2024. (Cited on pages 3 and 19.)

Thomas W. Malone and Michael S. Bernstein. *Handbook of Collective Intelligence*. MIT Press, 2015. (Cited on page 18.)

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *Causal Learning and Reasoning*, 2023. (Cited on page 8.)

Marvin Minsky. *The Society of Mind*. Simon and Schuster, 6. pb-pr. edition, 1988. (Cited on pages 2 and 18.)

Melanie Mitchell. *Complexity A Guided Tour*. Oxford University Press, 2009. (Cited on page 18.)

Thomas Nagel. Brain bisection and the unity of consciousness. *Synthese*, 22(3/4):396–413, 1971. (Cited on page 2.)

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000. (Cited on page 7.)

Caspar Oesterheld, Johannes Treutlein, Roger B Grosse, Vincent Conitzer, and Jakob Foerster. Similarity-based cooperative equilibrium. In *Advances in Neural Information Processing Systems*, 2023. (Cited on page 6.)

Samir Okasha. *Agents and Goals in Evolution*. Oxford University Press, 2018. (Cited on page 19.)

Laurent Orseau, Simon McGregor McGill, and Shane Legg. Agents and devices: A relative definition of agency. *arXiv*, 2018. arXiv:1805.12387. (Cited on pages 3 and 19.)

Elisabeth Pacherie. Intentional joint agency: Shared intention lite. *Synthese*, 190(10):1817–1839, 2013. (Cited on pages 3 and 18.)

Judea Pearl. *Causality*. Cambridge University Press, 2009. (Cited on pages 3 and 19.)

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017. (Cited on pages 5 and 6.)

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems*, 2024. (Cited on pages 2 and 20.)

Nina Rajcic and Anders Søgaard. Goal-directedness is in the eye of the beholder. *arXiv*, 2025. arXiv:2508.13247. (Cited on page 3.)

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007. (Cited on page 19.)

Debraj Ray. *A Game-Theoretic Perspective on Coalition Formation*. Oxford University Press, 2007. (Cited on pages 3 and 19.)

Alexander G. Reisach, Alberto Suárez, Sebastian Weichwald, and Antoine Chambaz. The case for time in causal DAGs. *arXiv*, 2025. arXiv:2501.19311. (Cited on page 30.)

Eigil F. Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Uncertainty in Artificial Intelligence*, 2021. (Cited on pages 8 and 12.)

Abraham Sesshu Roth. Shared Agency. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2017. (Cited on pages 1 and 18.)

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Uncertainty in Artificial Intelligence*, 2017. (Cited on pages 1, 3, and 19.)

David P. Schweikard and Hans Bernhard Schmid. Collective Intentionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2021. (Cited on page 1.)

John Searle. Collective intentions and actions. In Philip R. Cohen Jerry Morgan and Martha Pollack, editors, *Intentions in Communication*, pages 401–415. MIT Press, 1990. (Cited on pages 3 and 18.)

Anil Seth. Measuring emergence via nonlinear granger causality. In *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, 2006. (Cited on page 19.)

Joar Skalse and Alessandro Abate. Partial identifiability and misspecification in inverse reinforcement learning. *arXiv*, 2024. arXiv:2411.15951. (Cited on page 7.)

John Maynard Smith and Eörs Szathmáry. *The Major Transitions in Evolution*. Oxford University Press, 2020. Previously issued in print: Oxford : W.H. Freeman/Spektrum, 1995; Oxford: Oxford University Press, 1997. - Includes bibliographical references and index. - Description based on print version record and publisher information. (Cited on page 19.)

Oliver Sourbut, Lewis Hammond, and Harriet Wood. Cooperation and control in delegation games. In *International Joint Conference on Artificial Intelligence*, 2024. (Cited on page 19.)

Claudia Szabo and Yong Meng Teo. Formalization of weak emergence in multiagent systems. *ACM Transactions on Modeling and Computer Simulation*, 26(1):1–25, 2015. (Cited on page 19.)

Steven Tadelis. *Game theory: an introduction*. Princeton University Press, 2013. (Cited on page 27.)

Yong Meng Teo, Ba Linh Luong, and Claudia Szabo. Formalization of emergence in multi-agent systems. In *Proceedings of the 1st ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, 2013. (Cited on page 20.)

Raimo Tuomela. Joint intention, we-mode and I-mode. *Midwest Studies in Philosophy*, 30(1):35–58, 2006. (Cited on pages 3 and 18.)

Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. The reasons that agents act: Intention and instrumental goals. In *International Conference on Autonomous Agents and Multiagent Systems*, 2024. (Cited on page 19.)

Michael P. Wellman, Karl Tuyls, and Amy Greenwald. Empirical game theoretic analysis: A survey. *Journal of Artificial Intelligence Research*, 82, 2025. (Cited on page 30.)

Kevin Xia and Elias Bareinboim. Neural causal abstractions. In *AAAI Conference on Artificial Intelligence*, 2024. (Cited on page 6.)

Dylan Xu and Juan-Pablo Rivera. Towards measuring goal-directedness in AI systems. *arXiv*, 2024. arXiv:2410.04683. (Cited on pages 3, 19, and 31.)

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*, 2010. (Cited on page 19.)

Martin Zinkevich, Amy Greenwald, and Michael L. Littman. Cyclic equilibria in Markov games. In *Advances in Neural Information Processing Systems*, volume 18, pages 1641–1648, 2005. (Cited on page 30.)

Appendix A. Additional Related Work

Due to space constraints, our discussion of related work in Section 1.2 was rather brief. Here we extend this discussion along several axes.

Philosophical Foundations of Collective Agency. The question of when multiple individuals constitute a unified collective agent has long occupied philosophers. Early foundational work established the importance of collective intentionality – the capacity for shared intentions and joint commitments – in grounding group agency (Searle, 1990; Tuomela, 2006; Bratman, 2014). Subsequent work has refined these accounts, examining the semantics of collective intentional behavior (Ludwig, 2007), distinguishing between different modes of shared intention (Pacherie, 2013), and exploring the relationship between collective intentions and team agency (Gold and Sugden, 2007). List and Pettit (2011) provide a more functionalist account, arguing that a group can be considered an agent in its own right when it has representational states, motivational states, and a capacity to process these states, even when individual members may lack shared beliefs or desires. A key challenge across these accounts is the *aggregation problem*: determining when coordinated behavior among individuals reflects genuine collective agency rather than merely the sum of individual actions (Anderson, 1972; Roth, 2017). This question becomes particularly acute when considering systems that exhibit goal-directed behavior at multiple levels of abstraction, from neural subsystems within individual minds (Minsky, 1988) to emergent patterns in complex adaptive systems (Mitchell, 2009). It is worth noting that the question of collective *agency* is distinct from, though related to, that of collective *intelligence* (Malone and Bernstein, 2015), the latter focusing more on problem-solving capabilities than on unified agency per se. While these philosophical frameworks provide important conceptual foundations, they generally lack the formal machinery needed to rigorously identify collective agents or predict when they will emerge.

Causal Definitions of Individual Agency. Recent work has made important progress on rigorously defining and identifying *individual* agents using tools from causality and decision theory. Essentially, this is because any notion of goal-directedness – which is intrinsic to agency – requires considering what a putative agent would have done differently to achieve their goal, had their circumstances been different. Early work in AI and multi-agent systems developed formal models of joint action

and cooperation (Levesque et al., 1990), though these typically assumed rather than derived the conditions for collective agency. Most closely related to our work, Kenton et al. (2023) propose a causal discovery algorithm for identifying agents from empirical data. Others define and detect agency based on how predictive the assumption of utility-maximizing behavior is of a system's observed behavior: Orseau et al. (2018) use Bayesian inverse reinforcement learning (Ramachandran and Amir, 2007) to infer the extent to which a given system is a goal-directed agent, whereas MacDermott et al. (2024) use a maximal causal entropy model (Ziebart et al., 2010) to measure goal-directedness. More recently, others have attempted to apply such definitions to measure the goal-directedness not just of RL but also LLM-based agents (Xu and Rivera, 2024; Everitt et al., 2025). A related but separate research thread has studied causal definitions of intention (Halpern and Kleiman-Weiner, 2018; Ward et al., 2024). None of these works, however, formally consider the question of *collective* agents, where multiple subsystems have different actions, observations, and incentives.

Multi-Agent Formalisms and Collective Behavior. For settings with multiple agents, Hammond et al. (2023) provide a formal framework – known as causal games – by generalizing multi-agent influence diagrams (Koller and Milch, 2003) to the higher levels of Pearl's causal hierarchy (Pearl, 2009).[9] Other game-theoretic approaches, such as coalition structures in cooperative game theory (Ray, 2007; Elkind and Rothe, 2016), provide models of when subgroups might form and what outcomes they might achieve, but do not address whether such coalitions constitute unified agents. Complementing these formal frameworks, a substantial literature examines collective intelligence and emergent behavior in multi-agent systems, from swarm algorithms (Bonabeau et al., 1999) to collective decision-making in biological groups (Couzin, 2009). Several works have developed formal methods for detecting or measuring emergent phenomena by comparing micro- and macro-level descriptions (Kubík, 2003; Seth, 2006; Szabo and Teo, 2015) or by measuring changes in collective capabilities (Sourbut et al., 2024). Others have proposed theoretical frameworks for emergent agency drawing on active inference (Friston et al., 2022) or evolutionary perspectives (Okasha, 2018; Smith and Szathmáry, 2020). However, these approaches either measure emergence empirically without providing normative criteria for when collective agency exists, or apply to specific domains without offering a general framework for identifying collective agents across contexts.

Causal Abstraction. The notion that systems can be accurately described at multiple levels of abstraction has been formalized in the causal modelling literature through the concept of causal abstraction. Rubenstein et al. (2017) and Beckers and Halpern (2019) develop frameworks for determining when a high-level causal model is a valid abstraction of a more detailed low-level model, preserving causal relationships even when variables are aggregated or details are omitted. Beckers et al. (2020) extend this to allow for approximate abstractions that sacrifice some precision for greater simplicity. These frameworks provide rigorous criteria for when a simplified model faithfully represents the causal structure of a more complex system – precisely the kind of tool needed to determine when viewing a group of agents as a single collective agent is justified. However, existing causal abstraction work has focused primarily on single-agent causal influence diagrams or on causal models without decision-makers. Applying these ideas to multi-agent settings, where the low-level

9. In the same paper, Hammond et al. (2023) also introduce *mechanized* causal games, in which the parameters or decision rules of variables are explicitly represented as mechanism nodes, allowing Kenton et al. (2023) to discover which elements correspond to decisions and utilities.

model involves multiple decision-makers with potentially conflicting objectives, requires extending the abstraction framework to handle the strategic interactions captured by causal games.

AI Safety and Multi-Agent Risks. Our primary motivation is the safety and control of networks of advanced AI agents, and in particular the possibility of emergent 'super-agents' with dangerous or unexpected capabilities or goals. A key safety challenge is that multiple, simpler AI systems might inadvertently form a collective agent with capabilities and objectives distinct from any individual in the group (Hammond et al., 2025). Compared to AI tools, the ability of artificial *agents* to make plans and take actions in pursuit of complex goals makes them not only more useful, but also more harmful if those goals are misaligned (Chan et al., 2023). For example, competitive pressures may lead individually rational AI agents to rapidly exhaust collective resources (Piatti et al., 2024), or a group of agents might combine their harmless individual capabilities to override safeguards and execute dangerous attacks (Jones et al., 2024). Several works have examined emergent capabilities in multi-agent AI systems (Chen et al., 2009; Teo et al., 2013), though most focus on identifying specific emergent behaviors rather than general conditions for collective agency. Drexler (2019) proposes that a comprehensive set of narrow AI 'services' could provide a safer alternative to monolithic superintelligence, though this depends on preventing unwanted collective agents from emerging. Understanding when and how collective agency arises is therefore critical for anticipating and mitigating multi-agent AI risks.

Appendix B. Battle of the sexes example

Example 1 ***Battle of the sexes.*** *We illustrate the formalism with the classic two-player coordination game 'battle of the sexes', shown in Table 2.*

Table 2: Battle of the sexes payoff matrix

	Opera	Football
Opera	$(2,1)$	$(0,0)$
Football	$(0,0)$	$(1,2)$

The mechanized SCM $m\mathcal{M}$ consists of a deterministic cyclic SCM with signature

$$\begin{aligned}\widetilde{D}_1 &\in [0,1]\\ \widetilde{D}_2 &\in [0,1]\\ \widetilde{U}_1 &\in \mathbb{Z}^{\{O,F\}\times\{O,F\}}\\ \widetilde{U}_2 &\in \mathbb{Z}^{\{O,F\}\times\{O,F\}}.\end{aligned}$$

$\widetilde{D}_1$ and $\widetilde{D}_2$ specify probability distributions over 'opera' and 'football' (the number in $[0,1]$ represents the probability of picking opera). $\widetilde{U}_1$ and $\widetilde{U}_2 : \{O,F\} \times \{O,F\} \to \mathbb{Z}$ specify the payoffs of the two players' decisions (for example, in the battle of the sexes, $\widetilde{U}_1(O,O) = 2$, see Table 2). The

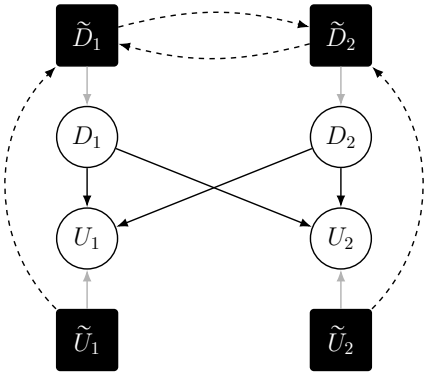

Figure 4: Mechanized causal graph for 2x2 games (of which battle of the sexes is an example). Dashed edges represent edges in the mechanisms model. Solid edges represent edges in the object-level model. Gray edges link mechanism variables with object variables. The distribution of D_1, which corresponds to the decision of player 1, depends on the conditional distribution of U_1 given D_1 and D_2 and the marginal distribution of D_2. This is represented by the dashed edges from $\widetilde{U}_1$ to $\widetilde{D}_1$ and $\widetilde{D}_2$ to $\widetilde{D}_1$, respectively. If player 1 is choosing a strategy using best response rationality (see Definition 10) with utility equal to the payoff, then player 1 will not respond to changes in $\widetilde{U}_2$ if $\widetilde{D}_2$ is kept fixed. Therefore, $\widetilde{D}_1$ does not functionally depend on $\widetilde{U}_2$ and we draw no dashed edge from $\widetilde{U}_2$ to $\widetilde{D}_1$.

parameterized SCM is given by the structural assignments

$$\mathcal{F}_{D_1}^{\widetilde{D}_1}(\mathcal{E}_{D_1}) = \begin{cases} O & \mathcal{E}_{D_1} \leqslant \widetilde{D}_1 \\ F & \textit{otherwise} \end{cases}$$
$$\mathcal{F}_{D_2}^{\widetilde{D}_2}(\mathcal{E}_{D_2}) = \begin{cases} O & \mathcal{E}_{D_2} \leqslant \widetilde{D}_2 \\ F & \textit{otherwise} \end{cases}$$
$$\mathcal{F}_{U_1}^{\widetilde{U}_1}(D_1, D_2, \mathcal{E}_{U_1}) = \widetilde{U}_1(D_1, D_2)$$
$$\mathcal{F}_{U_2}^{\widetilde{U}_2}(D_1, D_2, \mathcal{E}_{U_2}) = \widetilde{U}_2(D_1, D_2)$$

$\mathcal{E}_{D_1}, \mathcal{E}_{D_2} \sim \textit{Unif}(0, 1)$. *We assume that the payoffs are deterministic in the decisions, and so, the distributions of the noise variables* $\mathcal{E}_{U_1}, \mathcal{E}_{U_2}$ *are irrelevant.*

We now specify the mechanism model $\widetilde{\mathcal{M}}$. We assume that the payoffs are as specified in Table 2, that is,

$$\mathcal{F}_{\widetilde{U}_1}(\widetilde{\boldsymbol{V}}) = \begin{cases} (O,O) \mapsto 2 \\ (F,F) \mapsto 1 \\ (O,F) \mapsto 0 \\ (F,O) \mapsto 0 \end{cases}$$
$$\mathcal{F}_{\widetilde{U}_2}(\widetilde{\boldsymbol{V}}) = \begin{cases} (O,O) \mapsto 1 \\ (F,F) \mapsto 2 \\ (O,F) \mapsto 0 \\ (F,O) \mapsto 0. \end{cases}$$

The assignments for the payoff mechanisms constantly return the payoff function specified in Table 2 (irrespective of the policies of the players). We assume that the $\widetilde{D}_1$ and $\widetilde{D}_2$ act as to optimize the expected value of their respective payoffs, that is,

$$\mathcal{F}_{\widetilde{D}_1}(\widetilde{D}_2, \widetilde{U}_1, \widetilde{U}_2) \in \underset{\tilde{d}_1 \in [0,1]}{\arg\max}\, \mathbb{E}_{\widetilde{U}_1, \widetilde{U}_2, \tilde{d}_1, \widetilde{D}_2}(U_1)$$
$$\mathcal{F}_{\widetilde{D}_2}(\widetilde{D}_1, \widetilde{U}_1, \widetilde{U}_2) \in \underset{\tilde{d}_2 \in [0,1]}{\arg\max}\, \mathbb{E}_{\widetilde{U}_1, \widetilde{U}_2, \widetilde{D}_1, \tilde{d}_2}(U_2)$$

Notice that $\mathcal{F}_{\widetilde{D}_1}$ depends trivially on $\widetilde{U}_2$, and $\mathcal{F}_{\widetilde{D}_2}$ depends trivially on $\widetilde{U}_1$. It is well known that this system of equations has three solutions $\mathsf{S}(\widetilde{\mathcal{M}}) = \{\widetilde{s}_1, \widetilde{s}_2, \widetilde{s}_3\}$ where

$$\Pi_{\{\widetilde{D}_1, \widetilde{D}_2\}}(\widetilde{s}_1) = \{1_{\widetilde{D}_1}, 1_{\widetilde{D}_2}\}$$
$$\Pi_{\{\widetilde{D}_1, \widetilde{D}_2\}}(\widetilde{s}_2) = \{0_{\widetilde{D}_1}, 0_{\widetilde{D}_2}\}$$
$$\Pi_{\{\widetilde{D}_1, \widetilde{D}_2\}}(\widetilde{s}_3) = \left\{\frac{2}{3}_{\widetilde{D}_1}, \frac{1}{3}_{\widetilde{D}_2}\right\},$$

representing the three Nash equilibria.

Appendix C. First mover rationality

In this section, we define first mover rationality, which corresponds to making a decision and then assuming that the other players will adapt their decisions in response to that. In order to act like a first mover, an agent must have a belief model indicating which other nodes are believed to be agents and what their utility functions are believed to be.

Definition 18 *Let a mechanized causal graph $m\mathcal{M}$ and mechanism variable $\widetilde{S} \in \widetilde{\boldsymbol{V}}$ be given. A **belief model** is a tuple $(\widetilde{\boldsymbol{A}}, \boldsymbol{\mathcal{U}})$ where*

- *$\widetilde{\boldsymbol{A}} = (\widetilde{A}_1, \ldots, \widetilde{A}_n)$ is a sequence of distinct nodes in $\widetilde{\boldsymbol{V}} \backslash \{\widetilde{S}\}$. Intuitively, it corresponds to the set of nodes believed to be agents by $\widetilde{S}$.*

- $\boldsymbol{\mathcal{U}} = (\mathcal{U}_1, \ldots, \mathcal{U}_n)$ *is a sequence of utility functions. Intuitively,* $\mathcal{U}_i$ *is the utility believed by* $\widetilde{S}$ *to belong to agent* A_i*.*

Given a belief model, we can define first mover rationality as the decision that maximizes the expected utility given that other agents (according to the belief model) employ best response rationality.

Definition 19 *We define* ***first mover rationality*** $\mathcal{R}^{FM}_{\widetilde{S}}$ *for* $\widetilde{S} \in \widetilde{\boldsymbol{V}}$ *under belief model* $(\widetilde{\boldsymbol{A}}, \boldsymbol{\mathcal{U}})$ *and utility* $\mathcal{U}$ *by the condition that, for* $\widetilde{\mathbf{c}} \in \operatorname{ran}(\widetilde{\boldsymbol{V}} \backslash \{\widetilde{S}\})$*,*

$$\widetilde{s} \in \mathcal{R}^{FM}_{\widetilde{S}}(\widetilde{\mathbf{c}}) \text{ if and only if } \widetilde{s} \in \left\{ \Pi_{\widetilde{S}}(\widetilde{\boldsymbol{v}}) \mid \widetilde{\boldsymbol{v}} \in \arg\max_{\widetilde{\boldsymbol{w}} \in \boldsymbol{R}} \mathbb{E}_{\widetilde{\boldsymbol{w}}}(\mathcal{U}(\boldsymbol{V})) \right\},$$

where $\boldsymbol{R}$ *are those* $\widetilde{\boldsymbol{v}}$ *of* $\operatorname{ran}(\widetilde{\boldsymbol{V}})$ *for which*

- $\Pi_{\widetilde{\boldsymbol{V}} \backslash \widetilde{\boldsymbol{A}}}(\widetilde{\boldsymbol{v}}) = \Pi_{\widetilde{\boldsymbol{V}} \backslash \widetilde{\boldsymbol{A}}}(\widetilde{\mathbf{c}})$*. That is, non-agents do not respond by altering their distributions.*
- *For all* $\widetilde{A}_i \in \widetilde{\boldsymbol{A}}$*,* $\Pi_{\widetilde{A}_i}(\widetilde{\boldsymbol{v}}) \in \mathcal{R}^{BR}_{\widetilde{A}_i; \mathcal{U}_i}(\Pi_{\widetilde{\boldsymbol{V}} \backslash \{\widetilde{A}_i\}}(\widetilde{\boldsymbol{v}}))$*, where* $\mathcal{R}^{BR}_{\widetilde{A}_i; \mathcal{U}_i}$ *is the best response rationality relation for* $\widetilde{A}_i$ *with utility* $\mathcal{U}_i$*. That is, other agents are assumed to respond* $\mathcal{R}^{BR}$*-rationally to their own utility functions.*

In the following example we show that whether or not we have collective agency depends on the rationality principle employed by the agents. Concretely, we show an example where best response rationality does not induce collective agency, whereas first mover rationality does.

Example 2 ***Two agents with shared utility.***

Consider a situation where we have two $\mathcal{R}^{BR}$*-agents optimizing a shared utility function. Signature given by*

$$\begin{aligned} \widetilde{D}_1 &\in \{0, 1\} \\ \widetilde{D}_2 &\in \{0, 1\} \\ \widetilde{U} &\in \mathbb{R}^{\{0,1\} \times \{0,1\}}, \end{aligned}$$

and object variables

$$\begin{aligned} \mathcal{F}^{\widetilde{d}_1}_{D_1}(\mathcal{E}_{D_1}) &= \widetilde{d}_1 \\ \mathcal{F}^{\widetilde{d}_2}_{D_2}(\mathcal{E}_{D_2}) &= \widetilde{d}_2 \\ \mathcal{F}^{\widetilde{u}}_{U}(D_1, D_2, \mathcal{E}_U) &= \widetilde{u}(D_1, D_2). \end{aligned}$$

Assume that $\widetilde{D}_1$ *and* $\widetilde{D}_2$ *are both* $\mathcal{R}^{BR}$*-rational with respect to utility* $\mathcal{U}(\boldsymbol{v}) = \Pi_U(\boldsymbol{v})$*, that is,*

$$\begin{aligned} \mathcal{F}_{\widetilde{D}_1}(\widetilde{D}_2, \widetilde{U}) &\in \arg\max_{\widetilde{d}_1 \in \{0,1\}} \mathbb{E}_{\widetilde{U}, \widetilde{d}_1, \widetilde{D}_2}(U) \\ \mathcal{F}_{\widetilde{D}_2}(\widetilde{D}_1, \widetilde{U}) &\in \arg\max_{\widetilde{d}_2 \in \{0,1\}} \mathbb{E}_{\widetilde{U}, \widetilde{D}_1, \widetilde{d}_2}(U). \end{aligned}$$

Since both agents are optimizing the same utility function, we may wonder if we can abstract this model into a collective agent, that is, a model with a single ($\mathcal{R}^{BR}$, $\mathcal{U}$) agent. Consider high-level $m\mathcal{M}^*$ given by mechanism signature

$$\widetilde{D}^* \in \{0,1\}^2$$
$$\widetilde{U}^* \in \mathbb{R}^{\{0,1\}\times\{0,1\}}$$

and object variables

$$\mathcal{F}^{\widetilde{d}^*}_{D^*}(\mathcal{E}_{D^*}) = \widetilde{d}^*$$
$$\mathcal{F}^{\widetilde{u}^*}_{U^*}(D^*, \mathcal{E}_{U^*}) = \widetilde{u}^*(D^*).$$

Consider the value mappings

$$\tau_{D^*}(\{d_1, d_2\}) = (d_1, d_2)$$
$$\tau_{U^*}(u) = u$$

and intervention mappings given by

$$\omega_{\widetilde{D}^*}(\{\widetilde{d}_1, \widetilde{d}_2\}) = (\widetilde{d}_1, \widetilde{d}_2)$$
$$\omega_{\widetilde{U}^*}(\widetilde{u}) = \widetilde{u}.$$

Assume that $\widetilde{D}^*$ is $\mathcal{R}^{BR}$-rational with respect to utility U^* that is, $\mathcal{U}(\boldsymbol{v}^*) = \Pi_{U^*}(\boldsymbol{v}^*)$. Consider the following mechanism for U:

$$\widetilde{u} = (d_1, d_2) \mapsto \mathbb{1}(d_1 = d_2 = 0) + 2\mathbb{1}(d_1 = d_2 = 1)$$

We can convince ourselves that $m\mathcal{M}^*$ is not a mechanized abstraction of $m\mathcal{M}$ since $\mathsf{S}(\widetilde{\mathcal{M}}; \widetilde{u}) = \{\{0_{\widetilde{D}_1}, 0_{\widetilde{D}_2}, \widetilde{u}\}, \{1_{\widetilde{D}_1}, 1_{\widetilde{D}_2}, \widetilde{u}\}\}$ and $\mathsf{S}(\widetilde{\mathcal{M}}^*; \omega(u)) = \{\{(1,1)_{\widetilde{D}^*}, \widetilde{u}\}\}$, have different cardinalities.

If, on the other hand $\widetilde{D}_1$ and $\widetilde{D}_2$ had employed first mover rationality with accurate belief models, then

$$\mathcal{F}_{\widetilde{D}_1}(\widetilde{D}_2, \widetilde{U}) \in \arg\max_{\widetilde{d}_1 \in \{0,1\}} \max_{\widetilde{d}_2 \in \{0,1\}} \mathbb{E}_{\widetilde{U}, \widetilde{d}_1, \widetilde{d}_2}(U)$$
$$\mathcal{F}_{\widetilde{D}_2}(\widetilde{D}_1, \widetilde{U}) \in \arg\max_{\widetilde{d}_2 \in \{0,1\}} \max_{\widetilde{d}_1 \in \{0,1\}} \mathbb{E}_{\widetilde{U}, \widetilde{d}_1, \widetilde{d}_2}(U),$$

and $m\mathcal{M}^*$ would be a strong mechanized abstraction of $m\mathcal{M}$, suggesting that we can view $\widetilde{D}_1$ and $\widetilde{D}_2$ as a collective $\mathcal{R}^{BR}$-agent.

Appendix D. Formal description of the actor-critic example (Section 1.1 and Section 4.3)

Formally, consider the signature[10]

10. Rather than having object nodes W and Y, representing the utilities of critic and actor, respectively, we could have represented the utilities as external quantities computed from the object nodes, cf. Definition 8. Here, we choose to represent the utilities as object nodes to closely follow the presentation in Kenton et al. (2023).

$$
\begin{aligned}
\widetilde{R} &\in [0,1]^2 && \text{Probability of reward signal (+1) given state 0 or 1.}\\
\widetilde{W} &\in \{1\} && \text{Utility of the critic } W := -(R-Y)^2\\
\widetilde{Y} &\in \{1\} && \text{Utility of the actor } Y := Q(A)\\
\widetilde{S} &\in [0,1]^2 && \text{Probability of state 1 given action 0 or 1.}\\
\widetilde{Q} &\in [0,1]^2 && \text{Critic } (Q := \widetilde{Q})\\
\widetilde{A} &\in \{0,1\} && \text{Action } (A := \widetilde{A})
\end{aligned}
$$

And structural assignments encoding that the actor and critic are playing best response to Y and W, respectively:

$$
\begin{aligned}
\widetilde{R} &:= (r_0, r_1)\\
\widetilde{W} &:= 1\\
\widetilde{Y} &:= 1\\
\widetilde{S} &:= (s_0, s_1)\\
\widetilde{Q} &:= \Big(\widetilde{R}[0](1-\widetilde{S}[0]) + \widetilde{R}[1]\widetilde{S}[0], \widetilde{R}[0](1-\widetilde{S}[1]) + \widetilde{R}[1]\widetilde{S}[1]\Big)\\
\widetilde{A} &:= \mathbb{1}(\widetilde{Q}[0] \leqslant \widetilde{Q}[1]).
\end{aligned}
$$

Since the mechanism graph is acyclic, there is only one solution $\widetilde{\boldsymbol{s}}$ to this system of equations. Let us consider the following abstraction of the system:

$$
\begin{aligned}
\widetilde{R}^* &\in [0,1]^2 && \text{Probability of reward signal (+1) given state 0 or 1}\\
\widetilde{S}^* &\in [0,1]^2 && \text{Probability of state 1 given action}\\
\widetilde{A}^* &\in \{0,1\} && \text{Action } (A := \widetilde{A}^*),
\end{aligned}
$$

with structural assignments

$$
\begin{aligned}
\widetilde{R}^* &:= (r_0, r_1)\\
\widetilde{S}^* &:= (s_0, s_1)\\
\widetilde{A}^* &:= \mathbb{1}\left(\widetilde{S}^*[0]\widetilde{R}^*[1] + (1-\widetilde{S}^*[0])\widetilde{R}^*[0] \leqslant \widetilde{S}^*[1]\widetilde{R}^*[1] + (1-\widetilde{S}^*[1])\widetilde{R}^*[0]\right).
\end{aligned}
$$

To tie the models together, we define the value mappings and intervention mappings

$$
\begin{aligned}
\tau_{A*}(a) &= a & \qquad \omega_{\widetilde{A}*}(\widetilde{a}) &= \widetilde{a}\\
\tau_{S*}(s) &= s & \qquad \omega_{\widetilde{S}*}(\widetilde{s}) &= \widetilde{s}\\
\tau_{R*}(r) &= r & \qquad \omega_{\widetilde{R}*}(\widetilde{r}) &= \widetilde{r}.
\end{aligned}
$$

Appendix E. Proof of Proposition 16

Proposition 16. *Let mechanized models $m\mathcal{M}$ and $m\mathcal{M}^*$ be given. Assume that $m\mathcal{M}^*$ is a strong mechanized abstraction of $m\mathcal{M}$ for some value mappings $\{\tau_{V^*}\}_{V^*\in\boldsymbol{V}^*}$ and intervention mappings*

$\{\omega_{\widetilde{V}^*}\}_{\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*}$. *Let* $\widetilde{S}^* \in \widetilde{\boldsymbol{V}}^*$ *be a high-level mechanism variable. Assume that (i)* $\tau_{\mathbf{PA}_{S^*}}$ *is injective, and (ii) the mechanism nodes in* $\mathtt{A}_{\widetilde{S}^*}$ *have independent mechanisms. Then* $\widetilde{S}^*$ *is not a non-trivial agent in* m$\mathcal{M}^*$.

Proof Let $\widetilde{\boldsymbol{c}}_1^*, \widetilde{\boldsymbol{c}}_2^* \in \mathrm{ran}(\widetilde{\boldsymbol{V}}^* \backslash \{\widetilde{S}^*\})$ be arbitrary settings. Since m$\mathcal{M}^*$ is a strong abstraction of m$\mathcal{M}$, there exists low-level interventions $\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2 \in \mathrm{ran}(\mathtt{A}_{\widetilde{\boldsymbol{V}}^* \backslash \{\widetilde{S}^*\}})$ such that $\omega_{\widetilde{\boldsymbol{V}}^* \backslash \{\widetilde{S}^*\}}(\widetilde{\boldsymbol{y}}_1) = \widetilde{\boldsymbol{c}}_1^*$ and $\omega_{\widetilde{\boldsymbol{V}}^* \backslash \{\widetilde{S}^*\}}(\widetilde{\boldsymbol{y}}_2) = \widetilde{\boldsymbol{c}}_2^*$. Since the mechanism nodes in $\mathtt{A}_{\widetilde{S}^*}$ have independent mechanisms, $\mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}}_2)}(\mathtt{A}_{S^*} \mid \mathtt{A}_{\mathbf{PA}_{S^*}}) = \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}}_1)}(\mathtt{A}_{S^*} \mid \mathtt{A}_{\mathbf{PA}_{S^*}})$, which implies that $\mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}}_1)}(\tau_{S^*}(\mathtt{A}_{S^*}) \mid \tau_{\mathbf{PA}_{S^*}}(\mathtt{A}_{\mathbf{PA}_{S^*}})) = \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}}_2)}(\tau_{S^*}(\mathtt{A}_{S^*}) \mid \tau_{\mathbf{PA}_{S^*}}(\mathtt{A}_{\mathbf{PA}_{S^*}}))$ by injectivity of $\tau_{\mathbf{PA}_{S^*}}$. Causal consistency (Definition 14) now implies that

$$
\begin{aligned}
\mathbb{P}_{\widetilde{\boldsymbol{c}}_1^*, \mathcal{F}_{\widetilde{S}^*}(\widetilde{\boldsymbol{c}}_1^*)}(S^* \mid \mathbf{PA}_{S^*}) &= \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}}^*; \widetilde{\boldsymbol{c}}_1^*)}(S^* \mid \mathbf{PA}_{S^*}) \\
&= \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}}_1)}(\tau_{S^*}(\mathtt{A}_{S^*}) \mid \tau_{\mathbf{PA}_{S^*}}(\mathtt{A}_{\mathbf{PA}_{S^*}})) \\
&= \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}}_2)}(\tau_{S^*}(\mathtt{A}_{S^*}) \mid \tau_{\mathbf{PA}_{S^*}}(\mathtt{A}_{\mathbf{PA}_{S^*}})) \\
&= \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}}^*; \widetilde{\boldsymbol{c}}_2^*)}(S^* \mid \mathbf{PA}_{S^*}) \\
&= \mathbb{P}_{\widetilde{\boldsymbol{c}}_2^*, \mathcal{F}_{\widetilde{S}^*}(\widetilde{\boldsymbol{c}}_2^*)}(S^* \mid \mathbf{PA}_{S^*})
\end{aligned}
$$

Since $\widetilde{\boldsymbol{c}}_1^*, \widetilde{\boldsymbol{c}}_2^* \in \mathrm{ran}(\widetilde{\boldsymbol{V}}^* \backslash \{\widetilde{S}^*\})$ were arbitrary, we have that $\widetilde{S}^*$ is not a non-trivial agent. ■

Appendix F. Proof of Proposition 17

Proposition 17. *In Figure 2 (see also Appendix D for details),* m$\mathcal{M}^*$ *is a strong mechanized abstraction of* m$\mathcal{M}$. *Furthermore,* $\widetilde{A}^*$ *in the high-level model* m$\mathcal{M}^*$ *is an* $(\mathcal{R}^{BR}, \mathcal{U})$*-agent with utility equal to the reward* $\mathcal{U}(\boldsymbol{v}^*) = \Pi_{R^*}(\boldsymbol{v}^*)$.

Proof

To show that m$\mathcal{M}^*$ is a strong mechanized abstraction, we need to show that (1) consistency holds for all low-level interventions where ω is defined, and (2) $\omega_{\widetilde{V}^*}$ is surjective onto $\mathrm{ran}(\widetilde{V}^*)$ for all $\widetilde{V}^* \in \widetilde{\boldsymbol{V}}^*$.

First, consider surjectivity. Since the intervention mappings are the identity functions and the ranges of the relevant low-level and high-level variables are identical, we have surjectivity.

Second, we check consistency. Let $\widetilde{\boldsymbol{y}} = \{(s_0, s_1)_{\widetilde{S}}, (r_0, r_1)_{\widetilde{R}}\}$ be an arbitrary low-level intervention on $\mathtt{A}_{\widetilde{S}^*} \cup \mathtt{A}_{\widetilde{R}^*} = \{\widetilde{R}, \widetilde{S}\}$. Then $\omega(\widetilde{\boldsymbol{y}}) = \{(s_0, s_1)_{\widetilde{S}^*}, (r_0, r_1)_{\widetilde{R}^*}\}$ and

$$
\mathsf{S}(\widetilde{\mathcal{M}}^*; \omega(\widetilde{\boldsymbol{y}})) = \left\{\left\{(s_0, s_1)_{\widetilde{S}^*}, (r_0, r_1)_{\widetilde{R}^*}, \mathbb{1}\left(s_0 r_1 + (1 - s_0) r_0 \leqslant s_1 r_1 + (1 - s_1) r_0\right)_{\widetilde{A}^*}\right\}\right\}
$$

In the low-level model m$\mathcal{M}$:

$$
\begin{aligned}
\mathsf{S}(\widetilde{\mathcal{M}}; \widetilde{\boldsymbol{y}}) = \Big\{\Big\{&(r_0, r_1)_{\widetilde{R}}, 1_{\widetilde{Y}}, 1_{\widetilde{W}}, (s_0, s_1)_{\widetilde{S}}, \\
&(r_0(1 - s_0) + r_1 s_0, r_0(1 - s_1) + r_1 s_1)_{\widetilde{Q}}, \\
&\mathbb{1}\left(s_0 r_1 + (1 - s_0) r_0 \leqslant s_1 r_1 + (1 - s_1) r_0\right)_{\widetilde{A}}\Big\}\Big\}
\end{aligned}
$$

Since A^* and A are equal to 0 and 1 for the same values of r_0, r_1, s_0, s_1, we have that

$$\mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}};\widetilde{\boldsymbol{y}})}(\tau(\boldsymbol{V})) = \mathbb{P}_{\mathsf{S}(\widetilde{\mathcal{M}}^*;\omega(\widetilde{\boldsymbol{y}}))}(\boldsymbol{V}^*)$$

A similar calculation shows consistency for interventions on other elements in $\{\bigcup_{V^* \in \widetilde{\boldsymbol{Y}}^*} \mathtt{A}_{\widetilde{V}^*} \mid \widetilde{\boldsymbol{Y}}^* \subseteq \widetilde{\boldsymbol{V}}^*\}$.

We want to argue that in the abstracted model, $\widetilde{A}^*$ is $\mathcal{R}^{\text{BR}}$ rational under utility function $\mathcal{U}(\boldsymbol{v}^*) = \Pi_{R^*}(\boldsymbol{v}^*)$. So we must argue that $\widetilde{A}^*$ responds $\mathcal{R}^{\text{BR}}$-rationally to $\mathcal{U}$. Consider context $\widetilde{\boldsymbol{c}} = \{(s_0, s_1)_{\widetilde{S}^*}, (r_0, r_1)_{\widetilde{R}^*}\}$ and the responses $0_{\widetilde{A}^*}$ or $1_{\widetilde{A}^*}$:

$$\mathbb{E}_{\{0_{\widetilde{A}^*}, (s_0,s_1)_{\widetilde{S}^*}, (r_0,r_1)_{\widetilde{R}^*}\}}(R^*) = s_0 r_1 + (1-s_0) r_0$$
$$\mathbb{E}_{\{1_{\widetilde{A}^*}, (s_0,s_1)_{\widetilde{S}^*}, (r_0,r_1)_{\widetilde{R}^*}\}}(R^*) = s_1 r_1 + (1-s_1) r_0$$

So

$$\mathcal{R}^{\text{BR}}_{\widetilde{A}^*}(\widetilde{\boldsymbol{c}}) = \begin{cases} \{1_{\widetilde{A}^*}\} & s_1 r_1 + (1-s_1) r_0 > s_0 r_1 + (1-s_0) r_0 \\ \{0_{\widetilde{A}^*}\} & s_1 r_1 + (1-s_1) r_0 < s_0 r_1 + (1-s_0) r_0 \\ \{0_{\widetilde{A}^*}, 1_{\widetilde{A}^*}\} & s_1 r_1 + (1-s_1) r_0 = s_0 r_1 + (1-s_0) r_0, \end{cases}$$

and we conclude that $\mathcal{F}_{\widetilde{A}^*}(\widetilde{\boldsymbol{c}}) \in \mathcal{R}^{\text{BR}}_{\widetilde{A}^*}(\widetilde{\boldsymbol{c}})$ for every $\widetilde{\boldsymbol{c}} \in \text{ran}(\widetilde{\boldsymbol{V}}^* \backslash \{\widetilde{A}^*\})$. ■

Appendix G. Further details on Section 5

The code to reproduce all results is available at this link. It runs within a few minutes on a standard laptop.

G.1. Voting mechanisms

We consider three different voting mechanisms: Vickrey-Clarke-Groves (VCG), Median Voting, and Random Dictator.

VCG Voting. Under VCG voting, each citizen reports the parameters of their utility function and monetary transfers ensure that citizens are incentivized to report their true parameters (Tadelis, 2013): $\hat{a}_{ci} = (a_{ci} - \lambda_{ci}), \hat{b}_{ci} = b_{ci}, \hat{d}_{ci} = d_{ci}$. The mechanism then chooses the social optimum within each country, i.e.

$$q_c^{\text{NE}}(\boldsymbol{\lambda}) = \arg\max_{q_c \in \mathbb{R}} \sum_i \left(\hat{a}_{ci} q_c - \hat{b}_{ci} q_c^2 - \hat{d}_{ci} Q_W^2 \right).$$

While this voting mechanism is unrealistic, it serves as a simple baseline example: since this sum has the same form as Equation (2), we can calculate the low-level NE analytically using Equation (3). This also means that, for the VCG mechanism, we know that the high-level model is an exact mechanized abstraction of the low-level model, and we can compare the estimated parameters to ground-truth parameters.

Median Voting. For Median Voting, we cannot calculate the low-level NE analytically. Therefore, we approximate it using iterative best response. In particular, we initialize all pollution levels to 0 and then iteratively update the response of each country using the median vote within that country (using a damping factor of 0.3). We do this until convergence (tolerance of 1e-6 in the 2-norm).

Random Dictator. For Random Dictator voting, we randomly select a dictator from each country and use the dictators' utility functions to derive the low-level Nash equilibrium.

G.2. Further details on experiments

Population structure. We consider a population of 5 countries and 1000 citizens in total. The sizes are sampled from a log-normal distribution and scaled to sum to 1000. The countries and citizens are fixed throughout the experiments.

Citizen parameters. The citizen parameters are sampled i.i.d. uniformly within the ranges

$$\begin{aligned} a_{ci} &\in [0.35, 0.65], \\ b_{ci} &\in [7, 13]/N_c, \\ d_{ci} &\in [0.05, 0.15]/C, \end{aligned}$$

where N_c is the size of country c, and C is the number of countries. The parameter ranges are manually chosen to be somewhat reasonable.

Interventions λ. We sample different intervention means for each country. The means are sampled uniformly within the range $[0.0, 0.1]$. Given a mean λ_c, we sample interventions on individual citizens from the beta distribution with mean $10\lambda_c$ and concentration parameter 10. We then scale the sampled values back to the range $[0.0, 0.1]$.

δ_c estimation. For each country c, we sample 10 interventions where $(\lambda_{ci})_c = \mathbf{0}$, that is, no intervention on country c. We estimate δ_c by linear regression of q_c on Q_W (which is justified by Equation (3)). For random dictator voting, instead of estimating δ_c, we plug in the sum $\frac{1}{N_c}\sum_i \frac{d_{ci}}{b_{ci}}$ as an estimate (the mean of citizen-level $\delta_{ci} = d_{ci}/b_{ci}$). We do this to give the model a better chance of learning a useful ω. Estimating δ_c using the regression methods introduces bias since there is positive correlation between q_c and Q_W introduced by the stochasticity of picking the dictator. Alternatively, we could estimate δ_c using several regressions, each with a fixed dictator, and then take an average.

Surrogate model architecture. We use a fully connected neural network with 4 hidden layers of 128, 256, 256, and 128 neurons respectively. We use the ReLU activation function for all layers. The input is the intervention vector $(\lambda_{ci})_{ci}$ for all citizens and the output is α_c for each country.

Training procedure. We train the model on 1,000 intervention samples for 100 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 32. The results are evaluated on 500 held-out test interventions.

G.3. Further details on results

Country	Citizens	MAE(δ_c)	MAE(α_c)	MAE(q_c)	Baseline MAE(q_c)
0	180	0.000	0.025	0.011	0.233
1	107	0.000	0.029	0.007	0.159
2	202	0.000	0.065	0.017	0.256
3	412	0.000	0.107	0.022	0.468
4	99	0.000	0.025	0.005	0.138

Table 3: Results for VCG mechanism. We report the Mean Absolute Error (MAE) for the parameters δ_c and α_c, and for the pollution outcome q_c. We also include the baseline MAE for q_c, which is the MAE if we constantly predict the ground truth low-level Nash equilibrium at $\boldsymbol{\lambda} = \mathbf{0}$. The model is $\approx 95\%$ better than the baseline, see Table 1. For VCG we can recover δ_c perfectly using Equation (3) and two distinct $(q_c^{\mathrm{NE}}, Q_W^{\mathrm{NE}})$ pairs.

Country	Citizens	MAE(q_c)	Baseline MAE(q_c)
0	180	0.029	0.235
1	107	0.032	0.144
2	202	0.025	0.259
3	412	0.035	0.479
4	99	0.021	0.131

Table 4: Results for Median Voting mechanism. In contrast to VCG, we have no ground truth parameters α_c and δ_c to compare with. However, the surrogate model still accurately predicts the pollution outcome q_c, suggesting that a collective agency framework can effectively model median voting mechanisms.

Appendix H. Applications to MARL and LLM Agents

In this appendix, we discuss how our theoretical framework for collective agency may be applied to multi-agent reinforcement learning (MARL) and large language model (LLM) agents, and how it connects to AI safety. While a full treatment of these applications is beyond the scope of this work, we hope that the discussion below provides useful starting points for future research.

H.1. Applications to MARL

There are several subtleties that arise when attempting to translate insights from our framework in to the MARL setting. Here we provide a set of initial suggestions and observations that we hope makes the practical use of our formalism to MARL settings more straightforward. Such settings may provide a natural testbed for experimentation prior to developing real-world applications.

Approximate Equilibria and Bounded Rationality. A key feature of our framework is that it is agnostic to the specific notion of rationality employed by the agents. The rationality relation $\mathcal{R}$ (see Definition 10 and Appendix C) can encode a wide range of behavioral assumptions: exact best response, ε-best response, gradient-based learning rules, or any other decision procedure. This flexibility means that the framework applies even when agents are only approximately rational, as is typical in MARL, where agents may use gradient-based policy optimization and converge only to approximate equilibria. In such settings, one could define a rationality relation that captures

Country	Citizens	MAE(q_c)	Baseline MAE(q_c)
0	180	0.949	0.878
1	107	0.530	0.525
2	202	1.163	1.042
3	412	1.916	1.694
4	99	0.519	0.506

Table 5: Results for Random Dictator mechanism. We use $\delta_c = \frac{1}{N_c}\sum_i \frac{d_{ci}}{b_{ci}}$. The baseline is the average low-level random dictator Nash equilibrium at $\lambda = 0$ (averaged over dictator draws). The model performs worse than the baseline overall (-9.3%), see Table 1.

ε-optimality or local optimality, and then ask whether a collective agent model under this relaxed rationality relation constitutes a valid mechanized abstraction of the low-level system.

Learning Dynamics and Non-Stationarity. The mechanized causal games studied in this paper can be thought of as modeling equilibria: the solutions $\mathsf{S}(\widetilde{\mathcal{M}})$ correspond to settings of the mechanism variables that are mutually consistent given the rationality relations. On the one hand, this is useful because our framework applies regardless of the algorithm that agents use to reach an equilibrium (whether they use gradient-based learning, linear programming, or some other method). On the other hand, the models in this paper do not naturally support analysis of the learning dynamics themselves, or the non-stationarity that arises when agents are simultaneously updating their strategies. In particular, MARL algorithms can produce cycles in strategy space (Zinkevich et al., 2005), which are not captured by equilibrium-based models. Understanding the relationship between causal models and dynamical systems more generally is an active area of research; see, e.g., Reisach et al. (2025) for a discussion of complications and ambiguities related to the relationship between causal models and time, and Jørgensen et al. (2025) for ambiguities related to interventions.

Empirical Game-Theoretic Analysis. One promising avenue for applying our framework to large or complex MARL settings is through *empirical game-theoretic analysis* (Wellman et al., 2025). In this approach, overall policies (which might be large neural networks in practice) are modeled as pure strategies in a 'meta-game', which is then iteratively expanded and refined. By constructing a mechanized causal game at the level of the meta-game, it becomes possible to tractably analyze complex MARL settings using the tools from game theory and causal abstraction that we develop in this paper.

H.2. Applications to LLM-Based Multi-Agent Systems

Multi-agent systems composed of LLM-based agents are becoming increasingly common, from debate protocols (Irving et al., 2018; Du et al., 2024; Li et al., 2025) to multi-agent problem-solving and planning systems. Our framework provides a principled way to ask: when does a group of LLM agents, each with its own prompt, context, and objectives, act as if it were a single collective agent pursuing a coherent goal?

Collusion in Debate and AI Oversight. One possible application is to debate-based AI safety protocols (Irving et al., 2018). The effectiveness of AI safety via debate relies on the debating agents being adversarial rather than collectively pursuing a shared goal. If the debating agents were to collude – effectively forming a collective agent that optimizes for some objective other than truthful

argumentation – the safety guarantees of the debate protocol could break down. Similarly, in other settings where safety arises from some kind of automated oversight or adversarial relationship, cooperation is often undesirable. Our framework allows one to rule out collective agency under certain conditions (Proposition 16 provides a simple example), which could be used to provide assurance that mechanisms like debate are not illegibly pursuing unintended goals.

Emergent Goals in Networks of LLM Agents. More generally, whenever multiple LLM agents interact – whether through explicit communication channels, shared memory, or indirect coordination via a shared environment – our framework can be used to assess whether emergent collective agency is present and, if so, what objective the collective agent appears to be optimizing. Recent work on measuring goal-directedness in individual LLM agents (Xu and Rivera, 2024; Everitt et al., 2025) provides complementary tools that could be combined with our collective agency framework to detect and characterize emergent group-level goals in multi-agent LLM systems.