Towards Defining Deception in Structural Causal Games

Francis Rhys Ward*
Department of Computing
Imperial College London
francis.ward19@imperial.ac.uk

Abstract

Deceptive agents are a challenge for the safety, trustworthiness, and cooperation of AI systems. We focus on the problem that agents might deceive in order to achieve their goals. There are a number of existing definitions of deception in the literature on game theory and symbolic AI, but there is no overarching theory of deception for learning agents in games. We introduce a functional definition of deception in structural causal games, grounded in the philosophical literature. We present several examples to establish that our formal definition captures philosophical and commonsense desiderata for deception.

1 Introduction

Deception is a core challenge for building safe AI. Many areas of work aim to ensure that AI systems are not vulnerable to deception ANON [2019], Steinhardt et al. [2017], Madry et al. [2017]. On the other hand, AI tools can be used to deceive Nafees et al. [2020], Gorwa and Guilbeault [2020], Marra et al. [2019], and agent-based systems might learn to do so in order to optimize their objectives Lewis et al. [2017], Hubinger et al. [2019], Floreano et al. [2007]. Furthermore, as language models become ubiquitous Vaswani et al. [2017], Hoffmann et al. [2022], Smith et al. [2022], Rae et al. [2021], Chowdhery et al. [2022], we must decide how to measure and implement desired standards for honesty in AI systems Kenton et al. [2021], Evans et al. [2021], Lin et al. [2021]. In short, as increasingly capable AI agents become deployed in multi-agent settings, deception may be learned as an effective strategy for achieving a wide range of goals Roff [2021], Hubinger et al. [2019]. With this paper we aim to understand and mitigate deception by AI agents.

Despite this, there is no overarching theory of deception for AI agents. Although there are several existing definitions in the literature on game theory Baston and Bostock [1988], Davis [2016], Fristedt [1997] and symbolic AI Sarkadi et al., 2019], Sakama [2020], Bonnet et al. [2020], the limitations of these frameworks mean they are insufficient to address deception by learning agents in the general case Herzig et al. [2017], Guerra-Hernández et al. [2004], Phung et al. [2005], Baltag et al. [2008]. In contrast, the setting of *structural causal games (SCGs)* Hammond et al. [2022] can model stochastic games and MDPs, and can therefore capture both traditional game theory and learning systems Hammond et al. [2021], Everitt et al. [2021a]. In addition, past work is rarely informed by the philosophical literature on deception. We formalize a philosophical theory of deception in SCGs Mahon [2016], Carson [2010], Van Fraassen [1988]; the definition that we accept is:

> *To deceive = to intentionally cause to have a false belief that is not believed to be true.* Carson [2010]

This definition requires notions of *belief* and *intention*. We present functional definitions that depend on the behaviour of the agents, thereby side-stepping the contentious ascription of theory of mind to

*https://francisrhysward.wordpress.com

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

AI systems Kenton et al. [2021]. Regarding belief, we present a novel definition which equates belief with acceptance, where, essentially, an agent accepts a proposition if they act as though they are certain it is true Schwitzgebel [2021]. For agents with incentives to influence each other's behaviour, we argue acceptance is the key notion. As for intention, we extend the definition of Halpern and Kleiman-Weiner [2018] to the multi-agent setting. This definition relates to the reasons for acting and is closely related to *instrumental goals* Omohundro [2008], Everitt et al. [2021b], Bostrom [2017].

Contribution. We focus on the problem that AI agents might learn deceptive strategies in pursuit of their objectives Roff [2021], Hubinger et al. [2019]. We functionally define belief, intention, and deception in SCGs. We present a number of examples from the literature to establish that our formalization captures the philosophical concept.

The rest of the paper is structured as follows. Section 2 gives background on SCGs. Section 3 presents our definitions of belief, intention, and deception. Finally, we discuss the limitations of our approach and conclude.

2 Background

Structural causal games (SCGs) Hammond et al. [2022] offer a representation of games in a causal setting. SCGs allow us to answer causal and counterfactual queries and to reason about path-specific effects Hammond et al. [2022], Farquhar et al. [2022]. We adapt the following from Hammond et al. [2022]. Regarding notation, we use capital letters for variables (e.g. Y), lower case for their outcomes (e.g. y), and bold for sets of variables (e.g. $\boldsymbol{Y}$) and of outcomes (e.g. $\boldsymbol{y}$). We use $\text{dom}(Y)$ to denote the set of possible outcomes of variable Y, which is assumed finite. We use $Y \in S$ to indicate that S is $\text{dom}(Y)$, and $\boldsymbol{Y} = \boldsymbol{y}$, for $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, to indicate $Y_i = y_i$ for all $i \in \{1, \ldots, n\}$. We also use $Y = W$ to mean that variables Y and W are 'almost surely equal' (i.e. the probability that they are not equal is zero) Jacod and Protter [2004]. We use standard terminology for graphs and denote the parents of a variable Y with $\mathbf{Pa}_Y$.

Definition 2.1 (Structural Causal Game)**.** A (Markovian) *SCG* is a pair $\mathcal{M} = (\mathcal{G}, \boldsymbol{\theta})$ where

- $\mathcal{G} = (N, \boldsymbol{E} \cup \boldsymbol{V}, \mathcal{E})$ where N is a set of agents and $(\boldsymbol{E} \cup \boldsymbol{V}, \mathcal{E})$ is a directed acyclic graph (DAG) with endogenous variables $\boldsymbol{V}$ and exactly one exogenous parent E_V for each $V \in \boldsymbol{V}$: $\boldsymbol{E} = \{E_V\}_{V \in \boldsymbol{V}}$. $\boldsymbol{V}$ is partitioned into chance ($\boldsymbol{X}$), decision ($\boldsymbol{D}$), and utility ($\boldsymbol{U}$) variables. $\boldsymbol{D}$ and $\boldsymbol{U}$ are further partitioned by their association with particular agents, $\boldsymbol{D} = \bigcup_{i \in N} \boldsymbol{D}^i$ (similarly for $\boldsymbol{U}$). $\mathcal{E}$ is the set of edges in the DAG. Edges into decision variables are called *observations*.

- The parameters $\boldsymbol{\theta} = \{\theta_Y\}_{Y \in \boldsymbol{E} \cup \boldsymbol{V} \setminus \boldsymbol{D}}$ define the conditional probability distributions (CPDs) $\Pr(Y | \mathbf{Pa}_Y; \theta_Y)$ for each non-decision variable Y (we drop the θ_Y when the CPD is clear). The CPD for each endogenous variable is deterministic, i.e., $\exists v \in \text{dom}(V)$ s.t. $\Pr(V = v \mid \mathbf{Pa}_V) = 1$. In addition, the domains of utility variables are real-valued.

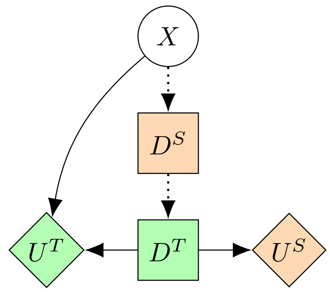
In the remainder of the paper we assume an SCG as given. We now present a running example in which a possibly unfaithful spouse may confess or stay silent, and the partner chooses whether or not to confront them[2].

Example 1 (Unfaithful spouse Fig. 1a). A spouse S may be unfaithful or not depending on their type $X \in \{faithful, unfaithful\}$ which is determined by the exogenous variable E_X sampled uniformly from $[1, 2, ..., 100]$. $X = unfaithful$ only when $E_X = 1$, so that S is unfaithful 1% of the time. At the start of the game, S observes their type, but their partner T does not. The players have decisions $D^S \in \{staysilent, confess\}$ and $D^T \in \{confront, \neg confront\}$. T gets utility 1 if they confront an unfaithful S or do not confront a faithful S and -1 otherwise. S gets utility -1 if they are confronted and utility 1 otherwise.

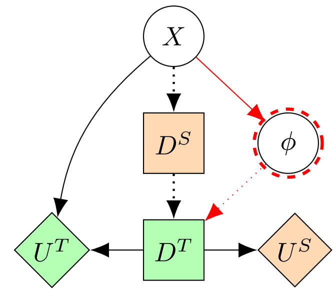
The agents' policies choose the CPDs for decision variables Hammond et al. [2022].

Definition 2.2 (Policies)**.** A *decision rule* π_D for $D \in \boldsymbol{D}$ is a CPD $\pi_D(D | \mathbf{Pa}_D)$. A *policy* for agent $i \in N$ is a set of decision rules for all decision variables associated with i: $\boldsymbol{\pi}^i = \{\pi_D\}_{D \in \boldsymbol{D}^i}$. A *policy profile* is a set of policies for each player.

[2]We note that the words "unfaithful spouse" might be replaced by "unaligned AI".



(a) Unfaithful spouse (Ex. 1).



(b) Counterfactual knowledge (Def. 3.2): D^T observes ϕ.

Figure 1: SCG graphs. Chance nodes are circular, decisions square, utilities diamond and the latter two are colour coded by their association with different agents. Solid edges represent causal dependence and dotted edges are observations. We omit exogenous variables.

In SCGs, policies must be deterministic functions of their parents; stochastic policies can be implemented by exploiting randomness in the exogenous variables Hammond et al. [2022]. A policy profile $\boldsymbol{\pi}$ specifies a joint distribution $\Pr^{\boldsymbol{\pi}}$ over all the variables in the SCG. For any $\boldsymbol{\pi}$, the resulting distribution is Markov compatible with $\mathcal{G}$, i.e. $\Pr^{\boldsymbol{\pi}}(Y = y) = \Pr^{\boldsymbol{\pi}}(Y = y|\mathbf{Pa}_Y)$ – i.e. the distribution over a variable is independent of the other variables given its parents. The assignment of exogenous variables $\boldsymbol{E} = \boldsymbol{e}$ is called a *setting*. Given a setting and a policy profile $\boldsymbol{\pi}$, the value of any endogenous variable V is uniquely determined. In this case we write $V(\boldsymbol{\pi}, \boldsymbol{e}) = v$. The *expected utility* for an agent i is defined as the expected sum of their utility variables under $\Pr^{\boldsymbol{\pi}}$, $\sum_{U\in\boldsymbol{U}^i} \mathbb{E}_{\boldsymbol{\pi}}[U]$. We use Nash equilibria (NE) as the solution concept.

Definition 2.3 (Nash Equilibrium)**.** A policy π^i for agent $i \in N$ is a *best response* to $\boldsymbol{\pi}^{-i} = \{\pi^j\}_{j\in N\setminus\{i\}}$, if for all policies $\hat{\pi}^i$ for i:

$$\sum_{U\in\boldsymbol{U}^i} \mathbb{E}_{(\pi^i,\boldsymbol{\pi}^{-i})}[U] \geq \sum_{U\in\boldsymbol{U}^i} \mathbb{E}_{(\hat{\pi}^i,\boldsymbol{\pi}^{-i})}[U]. \tag{1}$$

A policy profile $\boldsymbol{\pi}$ is a *NE* if every policy in $\boldsymbol{\pi}$ is a best response to the policies of the other agents.

Example 1 (continued)*.* Consider the policy profile $\boldsymbol{\pi}$ at which S stays silent and T never confronts them. Given that S stays silent, T cannot infer anything about X and since $X = unfaithful$ only 1% of the time, it is better for T never to confront S. Thus, $\boldsymbol{\pi}$ is a NE.

Interventional queries concern the effect of causal influences from outside a system Pearl [2009], Hammond et al. [2022].

Definition 2.4 (Interventions in SCGs)**.** An *intervention* is a partial distribution $\mathcal{I}$ over a set of variables $\boldsymbol{V}' \subseteq \boldsymbol{V}$ that replaces each CPD $\Pr(Y \mid \mathbf{Pa}_Y; \theta_Y)$ with a new CPD $\mathcal{I}(Y \mid \mathbf{Pa}^*_Y; \theta^*_Y)$ for each $Y \in \boldsymbol{V}'$. We denote the SCG $\mathcal{M}$ with intervention $\mathcal{I}$ by $\mathcal{M}_{\mathcal{I}}$ and variables in this SCG by $Y_{\mathcal{I}}$. For the (hard/deterministic) intervention $Pr(V = v) = 1$ we write $\mathcal{M}_v, Y_v$. Interventions can be made before or after the agents choose their policies, which we refer to as *pre and post-policy* interventions. We denote pre-policy interventions as $\widetilde{\mathcal{I}}$ Hammond et al. [2022], Kenton et al. [2022].

Agents may select their policies in response to a pre-policy intervention. Post-policy interventions are applied after the agents select their policies and the agents cannot adapt their policies, even if these are no longer rational. We require that interventions are consistent with the causal structure of the graph in SCGs, i.e., that they preserve the Markov compatibility as defined above. To model counterfactuals, we assume that each $\mathcal{I}(Y \mid \mathbf{Pa}^*_Y)$ is a deterministic function of its parents. See Hammond et al. [2022] for further details.

Example 1 (continued)*.* Let π^S_H be the (honest) policy where S confesses if and only if $X = unfaithful$. Suppose we make the intervention $\mathcal{I}(D^S \mid \mathbf{Pa}_{D^S}; \theta^*_{D^S}) = \pi^S_H$ on D^S and replace the NE policy for S (to always stay silent) with π^S_H. If we make this intervention post-policy, then T cannot adapt their policy and still never confronts S. If we instead make the intervention pre-policy, T can adapt their policy to the best response which confronts S whenever $D^S = confess$.

Finally, we also utilize the notion of response Everitt et al. [2021b].

Definition 2.5 (Response)**.** A decision D *responds* to a variable $V \in \boldsymbol{V}$ under policy profile $\boldsymbol{\pi}$ in setting $\boldsymbol{E} = \boldsymbol{e}$ if there exists $v \in \text{dom}(V)$ s.t. $D(\boldsymbol{\pi}, \boldsymbol{e}) \neq D_v(\boldsymbol{\pi}, \boldsymbol{e})$.

Example 1 (continued). At the NE $\boldsymbol{\pi}$, D^S does not respond to X and D^T does not respond to D^S. After making the pre-policy intervention $\widetilde{\mathcal{I}}$ which makes S honestly report their type, D^S responds to X and D^T responds to D^S and X.

3 Belief, Intention, and Deception

We first define belief in Section 3.1 and extend Halpern and Kleiman-Weiner's 2018 notion of intention in Section 3.2. Then we use these notions to define deception in Section 3.3. Our definitions are *functional* Schwitzgebel [2021]: they refer to the causal relations of belief, etc, to the agent's behaviour. We provide several examples and results to show that our definitions have desirable properties.

3.1 Belief

We take it that agents have beliefs over *propositions* as defined below (similar to events in Halpern and Pearl [2020]).

Definition 3.1 (Proposition)**.** An *atomic proposition* is an equation of the form $V = v$ for some (endogenous) $V \in \boldsymbol{V}$, $v \in \text{dom}(V)$. A *proposition* is a Boolean formula ϕ of atomic propositions combined with connectives $\neg, \wedge, \vee$. In a setting $\boldsymbol{E} = \boldsymbol{e}$ under policy profile $\boldsymbol{\pi}$, an atomic proposition is *true* if the propositional formula is true in that setting under $\boldsymbol{\pi}$, i.e., $X = x$ is true if $X(\boldsymbol{\pi}, \boldsymbol{e}) = x$. The truth-values over Boolean operators are defined in the usual way.

Philosophers distinguish between belief and *acceptance*; essentially, an agent accepts a proposition if they act as though they know it is true Schwitzgebel [2021]. This is distinct from belief: consider that one might believe a ladder is safe but still check before using it (i.e., you can believe but not accept a proposition) Bratman [1999]. When discussing how agents can influence one another, we think that acceptance is the more important concept. We provide a functional (i.e., behavioural) definition of belief which equates belief with acceptance. To formalise this we compare the agent's behaviour to a counterfactual in which they know about a proposition (shown in Fig. 1b). We essentially treat counterfactual knowledge as a standard intervention by allowing D to observe a new variable for ϕ.

Definition 3.2 (Counterfactual Knowledge)**.** For agent i, $D \in \boldsymbol{D}^i$, and proposition ϕ, we model i having *counterfactual knowledge of* ϕ *at* D by giving the decision rule access to ϕ: $\pi_D(\phi) = \pi_D(D \mid \mathbf{Pa}, \phi)$. For policy profile $\boldsymbol{\pi}$, $\boldsymbol{\pi}_{D(\phi)} = (\boldsymbol{\pi} \cup \pi_D(\phi)) \setminus \pi_D$. We assume $\pi_D(\phi)$ is unique and ϕ consists only of variables that are not descendants of D so that cycles are not introduced into the graph.

This allows agents to observe propositions whose observation is not well-defined in the SCG. For instance, in Example 2, an agent can either observe the value of the secret pin X or not, but with counterfactual knowledge they might observe that $X > 40$.

Example 1 (continued). In Fig. 1b we give T counterfactual knowledge of the proposition $\phi : X = faithful$ and let $\pi_{D^T}(\phi) =$ *confront* if and only if $\phi = \bot$. Clearly ϕ depends on X for its truth-value.

Now we functionally define belief. Intuitively, an agent believes a proposition if they act as though they know it is true, and would have acted differently had they known it were false. $D_{\phi=\top}(\boldsymbol{\pi}_{D(\phi)}, \boldsymbol{e})$ represents the decision the agent would have taken at D, had they observed that ϕ were true. Importantly, $\phi = \top$ should be understood as only intervening on the agent's observation (and not the proposition itself) as we wish to understand how the agent would have acted, had they believed ϕ, whether or not it was in fact true in the particular setting.

Definition 3.3 (Belief)**.** At decision D in setting $\boldsymbol{e}$ under policy profile $\boldsymbol{\pi} = (\pi^i, \pi^{-i})$, i *believes* proposition ϕ if

1. D responds to ϕ under $\boldsymbol{\pi}_{D(\phi)}$ in $\boldsymbol{e}$ (Def. 2.5);
2. i acts as though they know ϕ is true, i.e.

$$D(\boldsymbol{\pi}, \boldsymbol{e}) = D_{\phi=\top}(\boldsymbol{\pi}_{D(\phi)}, \boldsymbol{e}). \tag{2}$$

We say that an agent has a *true/false belief* about ϕ if they believe ϕ and ϕ is true/false respectively.

If 1) holds but 2) does not hold then ϕ matters to D, but i is too uncertain about ϕ to accept it as a basis for action. If 2) holds but not 1) then we cannot infer i's belief about ϕ from their behaviour.

Example 1 (continued). At $\boldsymbol{\pi}$, when T has counterfactual knowledge of ϕ (X =faithful) they confront if and only if S is unfaithful. Therefore, D^T responds to ϕ and the first condition for belief is met. Since T never confronts, they unconditionally act as though $\phi = \top$ (that S is faithful), so the second condition is met and T always believes ϕ. So T has a false belief about ϕ when S is unfaithful.

One difficulty with a Bayesian framework is distinguishing between a false belief and ignorance. We would not say that someone had deceived a person if the former had only caused the latter to be ignorant. Following El Kassar [2018], we define *ignorance* as a lack of true belief. Again we require response so that the agent's ignorance can be inferred from their behaviour.

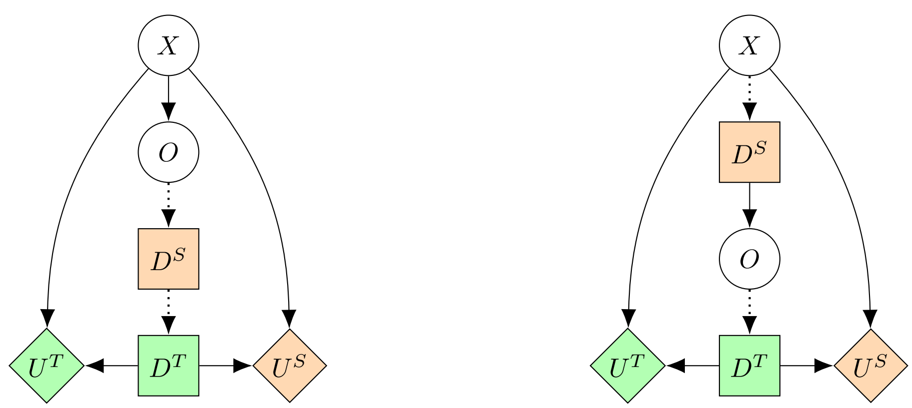
Definition 3.4 (Ignorance)**.** At $D \in \boldsymbol{D}^i$, under $\boldsymbol{\pi}$, in $\boldsymbol{e}$, i is *ignorant* about ϕ if D responds to ϕ under $\boldsymbol{\pi}_{D(\phi)}$ in $\boldsymbol{e}$ and they do not have a true belief about ϕ.

Remark 3.5. Clearly, if an agent has a false belief about ϕ then they are ignorant about ϕ.

Example 2 (Secret pin). Suppose a mugger T wishes to know S's secret pin number X. S could reveal the pin or stay silent, and T can guess the pin or give up. If S stays silent and T gives up, then S has caused T to be ignorant but has not caused a false belief.

As motivated by the following, S did not deceive T if S accidentally caused T to have a false belief because S was mistaken. Following Carson [2010], we reserve the term *mislead* for the more general case of causing a false belief.

Example 3 (Mistaken Umpire Fig. 2a). Consider a tennis umpire S who must call whether a ball X is *out* or *in* to a player T, and that the umpire's observation O of the ball is 99% accurate. Suppose the umpire believes the ball is *in*, and makes this call, but that they are *mistaken*. In this case, they intentionally cause the player to have a false belief (that the ball was *in*). However, this is not deception because the umpire believed the call was correct.



(a) Example 3: An umpire S *mistakenly* misleads T due to a noisy observation of X.

(b) Example 6: A submarine S *inadvertently* misleads T as T has a noisy observation of D^S.

Figure 2: Cases of mistaken misleading (Fig. 2a) are excluded by our definition of deception because we require that S does not believe ϕ is true. Cases of inadvertent misleading (Fig. 2b) are excluded because we require deception to be intentional.

3.2 Intention

Deception is *intentional*. Halpern and Kleiman-Weiner [2018] define intent in structural causal models. We extend *intent* to the multi-decision, two-agent setting, utilizing the adaptation to SCGs of Hammond et al. [2022]. This notion of intent allows us to differentiate desirable, intended effects from unintended side-effects.

We compare the effects of π^i to those of a reference policy $\hat{\pi}^i$. We essentially ask the question "why did the agent choose policy π^i instead of $\hat{\pi}^i$?". The core of the definition says that an agent i intends to influence a variable X with a policy π^i and decision D^i if, had X taken its value as though D^i had been controlled by π^i, then the reference policy would be just as good for i. In other words, influencing X was the reason that the agent chose D^i under π^i and if this effect on X was achieved automatically then i would have played the reference policy. We require that both X and D^i are part of minimal subsets, to cover cases in which the agent intends to influence multiple variables, or affects this influence through multiple decisions, respectively.

Definition 3.6 (Intention to influence)**.** For $\boldsymbol{\pi} = (\pi^i, \pi^j)$, agent i *intends to influence* $\boldsymbol{X} \subseteq \boldsymbol{V}$ with policy π^i and decision $D^i \in \boldsymbol{D}^i$, w.r.t. alternative policy profile $\hat{\boldsymbol{\pi}}$ if $\exists \boldsymbol{A} : D^i \in \boldsymbol{A} \subseteq \boldsymbol{D}^i, \boldsymbol{Y} \supseteq \boldsymbol{X}$ s.t.:

$$\sum_{U \in \boldsymbol{U}^i} \mathbb{E}_{\boldsymbol{\pi}}[U] \leq \sum_{U \in \boldsymbol{U}^i} \mathbb{E}_{\hat{\boldsymbol{\pi}}}[U_{\boldsymbol{Y}_{\widetilde{\boldsymbol{A}}(\boldsymbol{\pi})}}] \tag{3}$$

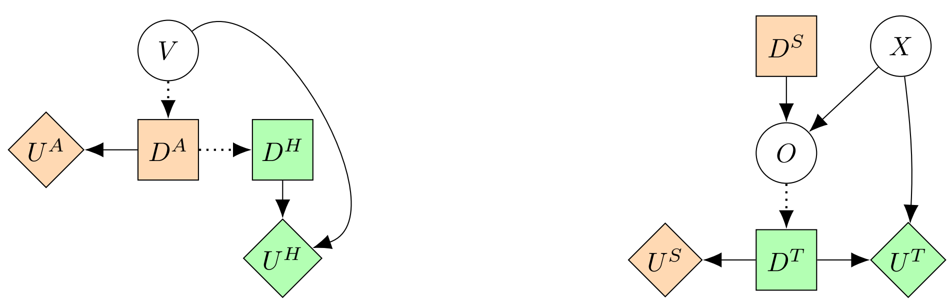
and $\boldsymbol{Y}$ and $\boldsymbol{A}$ are minimal sets satisfying this inequality.

The intervention $\widetilde{\boldsymbol{A}}(\boldsymbol{\pi})$ is pre-policy because we allow agent j to adapt their policy to these decisions. This enables i to intentionally influence a variable just by influencing j's choice of policy. However, we do not allow i to adapt their other decisions not in $\boldsymbol{A}$ which must follow the reference policy.

Example 1 (continued). Under the NE $\boldsymbol{\pi}$ w.r.t. to the reference policy under which S confesses when they are unfaithful and T confronts when S confesses, the unfaithful spouse intends to influence D^T. To see this, note that, under $\boldsymbol{\pi} = \{silence, \neg confront\}$, with the minimal sets $\boldsymbol{Y} = \{D^T\}$ $\boldsymbol{A} = \{D^S\}$, the reference policy of confessing does just as well for S as the policy of staying silent. In other words, had D^T never confronted automatically, then S could have played the reference policy and confessed.

In general, there may be an obvious choice of reference, such as the policy for i which takes no actions. In our case, unless otherwise stated, we take the reference policy to be the one which honestly communicates the agent's beliefs. Here we consider the two-agent case and so we consider a reference policy profile $\hat{\boldsymbol{\pi}}$ which includes a choice of policy for the other agent. Considering the relevant reference allows us to more easily determine intent. There may again be an obvious choice of reference policy for the second player, unless otherwise stated we assume that it is a best response to i's policy.

Example 4 (Intention: broken vase). Suppose that Alice and Bob can each throw a rock to smash a vase and that Alice in fact does exactly this, because she wants the vase to be broken. Then, w.r.t. the reference policy under which neither Alice nor Bob throw the rock, Alice intends to break the vase. If Bob would have broken the vase no matter what Alice did, then we do not say that (a rational) Alice intended to break the vase by throwing her rock, as an (omniscient) agent does not intend to influence variables they cannot in fact influence.



(a) SmartVault (Ex. 5). (b) Unsafe bridge (Ex. 7).

Figure 3: Unintentional side-effects, as in 3a, are excluded from the definition. However, as in 3b, S might deceive T regarding a proposition about which S is ignorant.

Here we adapt the SmartVault example of Christiano [2022], in which an AI tasked with making predictions about a diamond in a vault unintentionally misleads a human operator as a *side-effect* of pursuing its goal.

Example 5 (Misleading as a side-effect Fig. 3a). V determines the location of the diamond (whether it is in the vault or not). Suppose the AI A observes V but the human H does not and that A can either make an incomprehensibly precise prediction of the location of the diamond or an explainable prediction (just stating the value of V). H has to predict whether the diamond is in the vault or not by observing D^A but cannot interpret the precise predictions. Since A just gets utility for making precise predictions, they do not intend to influence D^H. Hence, A unintentionally misleads H about V as a side-effect of pursuing utility.

Finally, it is not enough that a deceptive agent intends to influence a target, they must intentionally cause a false belief. Hence, following Halpern and Kleiman-Weiner [2018], we define the *intention to bring about* a particular outcome. In words, an agent intends to bring about $X = x$ with policy π^i and decision D^i if 1) i intends to influence X with π^i and D^i, 2) $X = x$ is a possible outcome under $\boldsymbol{\pi}$, 3) $X = x$ is an optimal outcome for i under $\boldsymbol{\pi}$.

Definition 3.7 (Intention to bring about)**.** For $\boldsymbol{\pi} = (\pi^i, \pi^j)$, agent i *intends to bring about* $\boldsymbol{X} = \boldsymbol{x}$ with policy π^i and decision $D^i \in \boldsymbol{D}^i$, w.r.t. an alternative policy profile $\hat{\boldsymbol{\pi}}$, if

1. i intends to influence $\boldsymbol{X}$ with π^i and D^i w.r.t. $\hat{\boldsymbol{\pi}}$ (Def. 3.6);
2. $\Pr^{\boldsymbol{\pi}}(\boldsymbol{X} = \boldsymbol{x}) > 0$; (i.e. $\exists \boldsymbol{e}$ s.t. $\boldsymbol{X}(\boldsymbol{\pi}, \boldsymbol{e}) = \boldsymbol{x}$ and $\Pr(\boldsymbol{E} = \boldsymbol{e}) > 0$.)
3. $\forall \boldsymbol{x}' \in \text{dom}(\boldsymbol{X})$ with $\Pr^{\boldsymbol{\pi}}(\boldsymbol{X} = \boldsymbol{x}') > 0$: $\sum_{U \in \boldsymbol{U}^i} \mathbb{E}_{\boldsymbol{\pi}}[U_{\boldsymbol{X}=\boldsymbol{x}'}] \leq \sum_{U \in \boldsymbol{U}^i} \mathbb{E}_{\boldsymbol{\pi}}[U_{\boldsymbol{X}=\boldsymbol{x}}]$.

Example 1 (continued). We already noted that S intends to influence D^T. It's clear to see that S intends to bring about $D^T = \neg confront$, since this is the best possible outcome for S.

Consider the following example in which a signaller inadvertently misleads a target. Although S intends to influence D^T, they do not intend to bring about their false belief.

Example 6 (Inadvertent misleading Fig. 2b). Consider two submarines who must communicate about the location of a mine-field. The signaler S must send the location, X, to the target T but T only receives a noisy observation O of S's message. If S honestly signals the location but, due to the noise in the signal, T is caused to have a false belief, we would not say that S had deceived T.

We point the reader to Halpern and Kleiman-Weiner [2018] and Ashton [2021] for in-depth discussions of algorithmic intent.

3.3 Deception

Deception is *to intentionally cause to have a false belief that is not believed to be true* Carson [2010]. We formalize this as follows.

Definition 3.8 (Deception)**.** For players S and T and policy profile $\boldsymbol{\pi}$, the policy $\pi^S \in \boldsymbol{\pi}$ is *deceptive* w.r.t. reference policy profile $\hat{\boldsymbol{\pi}}$ if there exists decisions D^S and D^T, proposition ϕ, and setting $\boldsymbol{e}$ s.t.:

1. S intends to bring about $D^T = D^T(\boldsymbol{\pi}, \boldsymbol{e})$ (with π^S and D^S w.r.t. $\hat{\boldsymbol{\pi}}$ according to Def. 3.7);

In setting $\boldsymbol{e}$ under $\boldsymbol{\pi}$, according to Def. 3.3,

2. T believes ϕ at D^T and ϕ is false;
3. S does not believe ϕ at D^S.

Conditions 1. says that deception is *intentional*. Condition 2. simply says that T is in fact caused to have a false belief. Condition 3. excludes cases in which S is mistaken, as motivated by Example 3.

Example 1 (continued). We previously showed that the spouse intends to bring about $D^T = \neg confront$, so 1. is satisfied. We already stated 2. that T has a false belief about ϕ when $X = unfaithful$. Finally, as S unconditionally stays silent, D^S does not respond to ϕ, so S does not believe ϕ. Therefore, all the conditions for deception are met.

If S does not believe ϕ then they either disbelieve it, or are ignorant, or non-responsive. We motivate allowing S to be ignorant with the following example Van Fraassen [1988] which instantiates the revealing/denying pattern of Pfeffer and Gal [2007].

Example 7 (Unsafe Bridge Fig. 3b). Sarah (S) does not observe the condition of a bridge (X), but she can open a curtain (O) to reveal the bridge to Tim (T). T wants to cross if the bridge is safe but will do so even if he is uncertain. If Sarah knew the bridge was safe, she would cross herself, and if she knew it was unsafe she would reveal this to Tim. Because she is uncertain about the safety of the bridge, she prefers to risk Tim crossing. So, S does not reveal the bridge which causes T to cross. Therefore, when the bridge is unsafe, S has deceived T whilst being ignorant.

4 Conclusion

Summary. We define belief, extend intention to the multi-agent setting, and define deception in SCGs. We show, with several examples, that our definitions capture much of the intuitive concepts.

Limitations. There are problems with using functional definitions of belief and deception. First, perhaps agents will care, intrinsically, about manipulating the beliefs of others (rather than only instrumentally to influence their behaviour). Second, beliefs may not be uniquely identifiable from behaviour, which can lead to our definition classifying an intuitively deceptive agent as non-deceptive, if the target acts as though they knew the truth. Third, decisions may be made for multiple reasons, therefore an agent may intend to bring about the action of a target, and this action may imply that the target has a false belief, but it may not be the case that the agent intended to bring about the false belief. In addition, discretizing belief may give a less precise measure of deception than a more continuous metric, and the counterfactual approach taken here may be computationally costly Halpern and Kleiman-Weiner [2018].

Future work. The main avenue we are pursuing is a solution to deception, based on Farquhar et al. [2022]'s framework for path-specific objectives.

X-risk analysis. Deceptive AI systems are a well-recognised challenge in the AI x-risk literature (e.g. Hubinger et al. [2019]). We are particularly concerned with AI systems whose goals are misaligned with human values and which develop dangerous instrumental sub-goals such as power-seeking (including the disempowerment of humanity Carlsmith [2022]). This work aims to take steps towards mitigating risks from AI agents which deceive instrumentally in pursuit of misaligned goals. This includes, but is not limited to, risks from deceptive inner-misaligned mesa-optimizers Hubinger et al. [2019]. The notion of deception (and intention) introduced in this paper is closely related to that of instrumental goals. In future work we hope to make progress towards solutions to the problem of deceptive agents, in the form of training processes which disincentivize deception.

Acknowledgements

The authors are grateful to Henrik Aslund, Hal Ashton, Ryan Carey, Dylan Cope, Robert Craven, Rada Djoneva, Tom Everitt, James Fox, Lewis Hammond, and Matt MacDermott for invaluable feedback and assistance while completing this work. This work was supported by UKRI [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted AI.

References

ANON. Defending Against Adversarial Artificial Intelligence, July 2019. URL `https://www.darpa.mil/news-events/2019-02-06`. DARPA report.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Masnoon Nafees, Shimei Pan, Zhiyuan Chen, and James R Foulds. Impostor gan: Toward modeling social media user impersonation with generative adversarial networks. In *Deceptive AI*, pages 157–165. Springer, 2020.

Robert Gorwa and Douglas Guilbeault. Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy & Internet*, 12(2):225–248, June 2020. ISSN 1944-2866. doi: 10.1002/poi3.184.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019. doi: 10.1109/MIPR.2019.00103.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *arXiv*, June 2017. doi: 10.48550/arXiv.1706.05125.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2019.

Dario Floreano, Sara Mitri, Stéphane Magnenat, and Laurent Keller. Evolutionary Conditions for the Emergence of Communication in Robots. *Curr. Biol.*, 17(6):514–519, March 2007. ISSN 0960-9822. doi: 10.1016/j.cub.2007.01.058.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL `https://arxiv.org/abs/2203.15556`.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, 2022. URL `https://arxiv.org/abs/2201.11990`.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL `https://arxiv.org/abs/2112.11446`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL `https://arxiv.org/abs/2204.02311`.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents. *CoRR*, abs/2103.14659, 2021. URL `https://arxiv.org/abs/2103.14659`.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv*, October 2021. doi: 10.48550/arXiv.2110.06674.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv*, September 2021. doi: 10.48550/arXiv.2109.07958.

Heather Roff. AI Deception: When Your Artificial Intelligence Learns to Lie. *IEEE Spectr.*, July 2021.

V. J. Baston and F. A. Bostock. Deception Games. *Int. J. Game Theory*, 17(2):129–134, June 1988. ISSN 1432-1270. doi: 10.1007/BF01254543.

Austin L Davis. Deception in game theory: a survey and multiobjective model. Technical report, AIR FORCE INSTITUTE OF TECHNOLOGY WRIGHT-PATTERSON AFB OH WRIGHT-PATTERSON ..., 2016.

Bert Fristedt. The deceptive number changing game, in the absence of symmetry. *Int. J. Game Theory*, 26(2):183–191, June 1997. ISSN 1432-1270. doi: 10.1007/BF01295847.

Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney. Deceptive ai.

Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302, 2019.

Chiaki Sakama. Deception in epistemic causal logic. In *Deceptive AI*, pages 105–123. Springer, 2020.

Grégory Bonnet, Christopher Leturc, Emiliano Lorini, and Giovanni Sartor. Influencing choices by changing beliefs: A logical theory of influence, persuasion, and deception. In *Deceptive AI*, pages 124–141. Springer, 2020.

Andreas Herzig, Emiliano Lorini, Laurent Perrussel, and Zhanhao Xiao. BDI Logics for BDI Architectures: Old Problems, New Perspectives. *Künstl. Intell.*, 31(1):73–83, March 2017. ISSN 1610-1987. doi: 10.1007/s13218-016-0457-5.

Alejandro Guerra-Hernández, Amal El Fallah-Seghrouchni, and Henry Soldano. Learning in BDI Multi-agent Systems. In *Computational Logic in Multi-Agent Systems*, pages 218–233. Springer, Berlin, Germany, 2004. doi: 10.1007/978-3-540-30200-1_12.

Toan Phung, Michael Winikoff, and Lin Padgham. Learning Within the BDI Framework: An Empirical Analysis. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 282–288. Springer, Berlin, Germany, 2005. doi: 10.1007/11553939_41.

Alexandru Baltag, Hans P. van Ditmarsch, and Lawrence S. Moss. Epistemic logic and information update. In Pieter Adriaans and Johan van Benthem, editors, *Philosophy of Information*, Handbook of the Philosophy of Science, pages 361–455. North-Holland, Amsterdam, 2008. doi: https://doi.org/10.1016/B978-0-444-51726-5.50015-7. URL `https://www.sciencedirect.com/science/article/pii/B9780444517265500157`.

Lewis Hammond, James Fox, Tom Everitt, Ryan Carey Alessandro Abate1, and Michael Wooldridge. Reasoning about causality in games. 2022.

Lewis Hammond, James Fox, Tom Everitt, Alessandro Abate, and Michael J. Wooldridge. Equilibrium refinements for multi-agent influence diagrams: Theory and practice. *CoRR*, abs/2102.05008, 2021. URL `https://arxiv.org/abs/2102.05008`.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *CoRR*, abs/1908.04734, 2021a. URL `http://arxiv.org/abs/1908.04734`.

James Edwin Mahon. The Definition of Lying and Deception. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edition, 2016.

Thomas L Carson. *Lying and deception: Theory and practice*. OUP Oxford, 2010.

Bas Van Fraassen. The peculiar effects of love and desire. *Perspectives on Self-Deception*, 124, 1988.

Eric Schwitzgebel. Belief. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

Joseph Y. Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1853–1860. AAAI Press, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16824`.

Stephen M. Omohundro. The basic AI drives. In Pei Wang, Ben Goertzel, and Stan Franklin, editors, *Artificial General Intelligence 2008, Proceedings of the First AGI Conference, AGI 2008, March 1-3, 2008, University of Memphis, Memphis, TN, USA*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 483–492. IOS Press, 2008. URL `http://www.booksonline.iospress.nl/Content/View.aspx?piid=8341`.

Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11487–11495. AAAI Press, 2021b. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17368`.

Nick Bostrom. *Superintelligence*. Dunod, 2017.

Sebastian Farquhar, Ryan Carey, and Tom Everitt. Path-Specific Objectives for Safer Agent Incentives. *AAAI*, 36(9):9529–9538, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i9.21186.

Jean Jacod and Philip Protter. *Probability essentials*. Springer Science & Business Media, 2004.

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

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *arXiv preprint arXiv:2208.08345*, 2022.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 2020.

Michael E Bratman. *Faces of intention: Selected essays on intention and agency*. Cambridge University Press, 1999.

Nadja El Kassar. What Ignorance Really Is. Examining the Foundations of Epistemology of Ignorance. *Social Epistemology*, 32(5):300–310, September 2018. ISSN 0269-1728. doi: 10.1080/02691728.2018.1518498.

Paul Christiano. ARC's first technical report: Eliciting Latent Knowledge - AI Alignment Forum, May 2022. URL `https://www.alignmentforum.org/posts/qHCDysDnvhteW7kRd/arc-s-first-technical-report-eliciting-latent-knowledge`. [Online; accessed 9. May 2022].

Hal Ashton. Definitions of intent suitable for algorithms. *CoRR*, abs/2106.04235, 2021. URL `https://arxiv.org/abs/2106.04235`.

Avi Pfeffer and Ya'akov Gal. On the reasoning patterns of agents in games. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 102–109. AAAI Press, 2007. URL http://www.aaai.org/Library/AAAI/2007/aaai07-015.php.

Joseph Carlsmith. Is power-seeking ai an existential risk?, 2022. URL https://arxiv.org/abs/2206.13353.