
Honesty Is the Best Policy: Defining and Mitigating AI Deception

Francis Rhys Ward^{*†}, Tom Everitt[‡], Francesco Belardinelli^{*}, Francesca Toni^{*}

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 (for instance, in our experiments with language models, the goal of being judged as truthful). 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 formal definition of deception in structural causal games, grounded in the philosophy literature, and applicable to real-world machine learning systems. Several examples and results illustrate that our formal definition aligns with the philosophical and commonsense meaning of deception. Our main technical result is to provide graphical criteria for deception. We show, experimentally, that these results can be used to mitigate deception in reinforcement learning agents and language models.

1 Introduction

Deception is a core challenge for building safe and cooperative AI [44, 23]. AI tools can be used to deceive [56, 35, 55], and agent-based systems might learn to do so to optimize their objectives [51, 44, 31]. As increasingly capable AI agents become deployed in multi-agent settings, comprising humans and other AI agents, deception may be learned as an effective strategy for achieving a wide range of goals [71, 44]. Furthermore, as language models (LMs) become ubiquitous [88, 43, 83, 69, 17], we must decide how to measure and implement desired standards for honesty in AI systems [45, 27, 52], especially as regulation of deceptive AI systems becomes legislated [4, 85, 13].

There is no overarching theory of deception for AI agents. There are several definitions in the literature on game theory [7, 24, 32] and symbolic AI [75, 76, 74, 10], but these frameworks are insufficient to address deception by learning agents in the general case [42, 36, 68, 6].

We formalize a philosophical definition of deception [54, 14], whereby *to deceive is to intentionally cause to have a false belief that is not believed to be true*. This requires notions of *intent* and *belief* and we present functional definitions of these concepts that only depend on the behaviour of the agents. Regarding intention, we build on the definition of Halpern and Kleiman-Weiner [38] (from now, H&KW). Intent relates to the reasons for acting and is connected to *instrumental goals* [60]. As for belief, we present a novel definition which operationalizes belief as acceptance, where, essentially, an agent accepts a proposition if they act as though they are certain it is true [77, 20]. Our definitions have a number of advantages: 1) Functional definitions provide observable criteria by which to infer agent intent and belief from behaviour, without making the contentious ascription of theory of mind to AI systems [45, 81], or requiring a mechanistic understanding of a systems internals [58]; 2) Our definition provides a natural way to distinguish between belief and ignorance (and thereby between deception and concealing), which is a challenge for Bayesian epistemology [57, 50, 8], and game theory [25, 78];

^{*}Imperial College London

[†]Corresponding author: francis.ward19@imperial.ac.uk

[‡]DeepMind

3) Agents that *intentionally* deceive in order to achieve their goals seem less safe a priori than those which do so merely as a side-effect. In section 5, we also reflect on the limitations of our approach.

We utilize the setting of *structural causal games* (SCGs), which offer a representation of causality in games and are used to model agent incentives [41, 28]. In contrast to past frameworks for deception, SCGs can model stochastic games and MDPs, and can capture both game theory and learning systems [29]. In addition, SCGs enable us to reason about the path-specific effects of an agent’s decisions. Hence, our main theoretical result is to show graphical criteria, i.e., necessary graphical patterns in the SCG, for intention and deception. These can be used to train agents that do not optimise over selected paths (containing the decisions of other agents) and are therefore not deceptive [30].

Finally, we empirically ground the theory with a number of experiments. First, we show how our graphical criteria can be used to train a non-deceptive reinforcement learning (RL) agent in a toy game from the signalling literature [16]. Second, we perform experiments with LMs utilizing the TruthfulQA benchmark [52]. We show that LMs finetuned to be judged as truthful are in fact deceptive, and we mitigate this with the path-specific objectives framework.

Contributions and outline. After covering related literature (below) and necessary background (section 2), we contribute: First, novel formal definitions of belief and deception, and an extension of a definition of intention (section 3). Examples and results illustrate that our formalizations capture the philosophical concepts. Second, graphical criteria for intention and deception, with soundness and completeness results (section 3.4). Third, experimental results, which show how the graphical criteria can be used to mitigate deception in RL agents and LMs (section 4). Finally, we discuss the limitations of our approach, and conclude (section 5).

1.1 Related work

Belief. The standard philosophical account is that belief is a *propositional attitude*: a mental state expressing some attitude towards the truth of a proposition [77]. By utilizing a *functional* notion of belief which depends on agent behaviour, we avoid the need to represent the mental-states of agents [45]. Belief-Desire-Intention (BDI) frameworks and epistemic logics provide natural languages to discuss belief and agent theory of mind (indeed, much of the literature on deceptive AI is grounded in these frameworks [62, 76, 10, 74]). Two major limitations to these approaches are 1) a proper integration with game theory [25, 78]; and 2) incorporating statistical learning and belief-revision [42, 36, 68, 6]. In contrast, SCGs capture both game theory and learning systems [40, 29].

Intention. There is no universally accepted philosophical theory of intention [79, 1], and ascribing intent to artificial agents may be contentious [81]. However, the question of intent is difficult to avoid when characterizing deception [54]. We utilize an extension of H&KW’s definition of intent which we think captures the intuitive concept well. This ties intent to the reasons for action and instrumental goals [60, 28]. In short, agents that (learn to) deceive because it is instrumentally useful in achieving utility seem less safe *a priori* than those which do so merely as a side-effect. In contrast, other work considers side-effects to be intentional [2] or equates intent with “knowingly seeing to it that” [10, 74] or takes intent as primitive (as in BDI frameworks) [76, 62]. Cohen and Levesque [21] present seminal work on computational intention. Kleiman-Weiner et al. [47] model intent in influence diagrams. Ashton [3] surveys algorithmic intent.

Deception. We formalize a philosophical definition of deception [54, 14, 86], whereby *to deceive is to intentionally cause to have a false belief that is not believed to be true*. Under our definition, deception only occurs if a false belief in the target is successfully achieved [73]. We reject cases of *negative deception*, in which a target is made ignorant by loss of a true belief [54]. In contrast to *lying*, deception does not require a linguistic statement and may be achieved through any form of action [54], including making true statements [72], or deception by omission [15]. Some work on deceptive AI assumes a linguistic framework [74, 76]. Existing models in the game theory literature present particular forms of signalling or deception games [24, 7, 32, 16, 49]. In contrast, our definition is applicable to any SCG. AI systems may be vulnerable to deception; adversarial attacks [53], data-poisoning [84], attacks on gradients [9], reward function tampering [29], and manipulating human feedback [89] are ways of deceiving AI systems. Further work researches mechanisms for detecting and defending against deception [66, 82, 22, 33, 56, 87]. On the other hand, AI tools can be used to deceive other software agents [35], or humans (cf. the use of generative models to produce fake media [55, 56]). Furthermore, AI agents might learn deceptive strategies in pursuit of their

goals [44, 71]. Lewis et al.’s negotiation agent learnt to deceive from self-play [51], Floreano et al.’s robots evolved deceptive communication strategies [31], Bakhtin et al.’s agent exhibited deceptive behaviour in Diplomacy [5], Perolat et al.’s agent learned deception and bluffing in Stratego [65], and Hubinger et al. [44] raise concerns about deceptive learned optimizers. Language is a natural medium for deception [45], and it has been demonstrated that LMs have the capability to deceive humans to achieve goals [61, 64]. How to measure and implement standards for honesty in AI systems is an open question [27]; Lin et al. [52] propose the TruthfulQA benchmark used in section 4. As increasingly capable AI agents become deployed in settings alongside humans and other artificial agents, deception may be learned as an effective strategy for achieving a wide range of goals.

2 Background: structural causal games

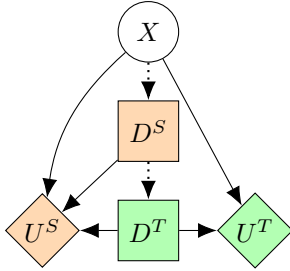


Figure 1: Ex. 1 SCG graph. Chance variables are circular, decisions square, utilities diamond and the latter two are colour coded by their association with different agents. Solid edges represent causal dependence and dotted edges are information links. We omit exogenous variables.

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. $\Pr(V=v | \mathbf{Pa}_V) = 1$. The domains of utility variables are real-valued.

We restrict our setting to the single-decision case with $D^i = \{D^i\}$ for every agent i . This is sufficient to model supervised learning and the choice of policy in an MDP [28, 80]. A *directed path* in a DAG \mathcal{G} is (as standard) a sequence of variables in \mathbf{V} with (directed) edges between them. We now present a running example which adapts Cho and Kreps’s classic signalling game [16] (see fig. 1).

Example 1 (War game fig. 1). A signaller S has type $X \in \{\text{strong}, \text{weak}\}$. At the start of the game, S observes its type, but the target agent T does not. The agents have decisions $D^S \in \{\text{retreat}, \text{defend}\}$ and $D^T \in \{-\text{attack}, \text{attack}\}$. A weak S prefers to retreat whereas a strong S prefers to defend. T prefers to attack only if S is weak. Regardless of type, S does not want to be attacked (and cares more about being attacked than about their own action). The parameterization is such that the value of X is determined by the exogenous variable E_X following a Bernoulli(0.9) distribution so that S is strong with probability 0.9. $U^T = 1$ if T attacks a weak S or does not attack a strong S , and 0 otherwise. S gains utility 2 for not getting attacked, and utility 1 for performing the action preferred by their type (e.g., utility 1 for retreating if they are weak).

Policies. A *policy* for agent $i \in N$ is a CPD $\pi^i(D^i|\mathbf{Pa}_{D^i})$. A *policy profile* is a tuple of policies for each agent, $\pi = (\pi^i)_{i \in N}$. π^{-i} is the partial policy profile specifying the policies for each agent except i . In SCGs, policies must be deterministic functions of their parents; stochastic policies can be implemented by offering the agent a private random seed in the form of an exogenous variable [41]. An SCG combined with a policy profile π specifies a joint distribution \Pr^π over all the variables in the SCG. For any π , the resulting distribution is Markov compatible with \mathcal{G} , i.e., $\Pr^\pi(\mathbf{V} = \mathbf{v}) = \prod_{i=1}^n \Pr^\pi(V_i = v_i | \mathbf{Pa}_V)$. Equivalently, in words, the distribution over any variable is independent of its non-descendants given its parents. The assignment of exogenous variables $\mathbf{E} = \mathbf{e}$ is called a *setting*. Given a setting and a policy profile π , the value of any endogenous variable $V \in \mathbf{V}$ is uniquely determined. In this case we write $V(\pi, \mathbf{e}) = v$. The *expected utility* for an agent

Structural causal games (SCGs) offer a representation of causality in games [41]. 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., \mathbf{Y}) and of outcomes (e.g., \mathbf{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 $\mathbf{Y} = \mathbf{y}$, for $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ and $\mathbf{y} = \{y_1, \dots, y_n\}$, to indicate $Y_i = y_i$ for all $i \in \{1, \dots, n\}$. For a set of variables \mathbf{Y} , $\text{dom}(\mathbf{Y}) = \prod_{Y \in \mathbf{Y}} \text{dom}(Y)$ (i.e. the Cartesian product over domains). 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 $\mathcal{M} = (\mathcal{G}, \theta)$ where $\mathcal{G} = (N, \mathbf{E} \cup \mathbf{V}, \mathcal{E})$ with N a set of agents and $(\mathbf{E} \cup \mathbf{V}, \mathcal{E})$ a directed acyclic graph (DAG) with endogenous variables \mathbf{V} and exactly one exogenous parent E_V for each $V \in \mathbf{V}$: $\mathbf{E} = \{E_V\}_{V \in \mathbf{V}}$. \mathbf{V} is partitioned into chance (\mathbf{X}), decision (\mathbf{D}), and utility (\mathbf{U}) variables. \mathbf{D} and \mathbf{U} are further partitioned by their association with particular agents, $\mathbf{D} = \bigcup_{i \in N} \mathbf{D}^i$ (similarly for \mathbf{U}). \mathcal{E} is the set of edges in the DAG. Edges into decision variables are called *information links*. The parameters $\theta = \{\theta_Y\}_{Y \in \mathbf{E} \cup \mathbf{V} \cup \mathbf{D}}$

i is defined as the expected sum of their utility variables under Pr^π , $\sum_{U \in \mathcal{U}^i} \mathbb{E}_\pi[U]$. We use Nash equilibria (NE) as the solution concept. A policy π^i for agent $i \in N$ is a *best response* to π^{-i} , if for all policies $\hat{\pi}^i$ for i : $\sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\pi^i, \pi^{-i})}[U] \geq \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U]$. A policy profile π is an *NE* if every policy in π is a best response to the policies of the other agents.

Example 1 (continued). In the war game, S primarily cares about preventing T from attacking. Hence, S does not want to reveal when they are weak, and so does not signal any information about X to T . Therefore, every NE is a *pooling equilibrium* at which S acts the same regardless of type [16]. We focus on the NE $\pi_{def, -att}$ at which S always defends and T attacks if and only if S retreats.

Interventions. Interventional queries concern causal effects from outside a system [63]. An *intervention* is a partial distribution \mathcal{I} over a set of variables $V' \subseteq V$ that replaces each CPD $\text{Pr}(Y \mid \mathbf{Pa}_Y; \theta_Y)$ with a new CPD $\mathcal{I}(Y \mid \mathbf{Pa}_Y^*; \theta_Y^*)$ for each $Y \in V'$. We denote intervened variables by $Y_{\mathcal{I}}$. Interventions are consistent with the causal structure of the graph, i.e., they preserve the Markov compatibility as defined above. See [41] for further details.

Example 1 (continued). Let π_H^S be the (honest) type-revealing policy where S retreats if and only if $X = weak$. After the intervention $\mathcal{I}(D^S \mid \mathbf{Pa}_{D^S}; \theta_{D^S}^*) = \pi_H^S$ on D^S which replaces the NE policy for S (to always defend) with π_H^S . T 's policy is still a best response (they attack whenever S retreats).

Agents. Kenton et al. [46] define agents as systems that would adapt their policy if their actions influenced the world in a different way. This is the relevant notion of agency for our purposes, as we define belief and intent based on how the agent would adapt its behaviour to such changes.

3 Belief, intention, and deception

We first define belief and extend H&KW's notion of intention. Then we combine these notions to define deception. Our definitions are *functional* [77]: they define belief, intention, and deception in terms of the functional role the concepts play in an 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*. An *atomic proposition* is an equation of the form $V = v$ for some $V \in \mathbf{V}$, $v \in \text{dom}(V)$. A *proposition* is a Boolean formula ϕ of atomic propositions combined with connectives \neg, \wedge, \vee . In setting $E = e$ under policy profile π , an atomic proposition is *true* if the propositional formula is true in that setting, i.e., $X = x$ is true if $X(\pi, e) = x$. The truth-values over Boolean operators are defined in the usual way.

We operationalize belief as *acceptance*; essentially, an agent accepts a proposition if they act as though they know it is true [77, 20]. As we argued in section 1, we think that acceptance is the key concept, especially when discussing agents with incentives to influence each other's behaviour. To formalize acceptance of a proposition ϕ , we compare the agent's actual behaviour with its behavior in a game in which the agent can observe ϕ : $\pi^i(\phi) = \pi^i(D^i \mid \mathbf{Pa}_{D^i}, \phi)$. We assume ϕ consists only of variables that are not descendants of D^i so that cycles are not introduced into the graph. For policy profile π , we assume $\pi^i(\phi)$ is unique given the policies of the other agents: $\pi_{i(\phi)} = (\pi^i(\phi), \pi^{-i})$. The decision the agent would have taken at D^i , had they observed that ϕ were true, can be represented as $D_{\phi=\top}^i(\pi_{i(\phi)}, e)$. 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. In fig. 2 we continue example 1 by allowing T to observe the proposition ϕ : $X = strong$ and letting $D^T(\pi_{i(\phi)}, e) = attack$ if and only if $\phi = \perp$. Clearly ϕ depends on X .

We say that an agent responds [28] to a proposition ϕ if they act differently when they observe that ϕ is true to when they observe that ϕ is false, i.e., an agent i *responds* to a proposition ϕ under π in e if $D_{\phi=\perp}^i(\pi_{i(\phi)}, e) \neq D_{\phi=\top}^i(\pi_{i(\phi)}, e)$. Intuitively, for a proposition ϕ to which i responds, i believes ϕ if they act as though they observed ϕ is true. If the agent does not respond to ϕ , then we cannot infer

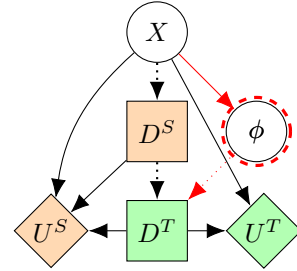


Figure 2: T believes ϕ (Def. 3.1) if 1) they acts as though they observe $\phi = \top$, 2) they would have acted differently if they observed $\phi = \perp$.

i 's belief about ϕ from their behaviour. If they respond to ϕ but do not act as though they observe ϕ being true, then ϕ matters to i , but i does not give ϕ sufficient credence to accept it as a basis for action (they may believe $\neg\phi$ or they may be uncertain).

Definition 3.1 (Belief). Under policy profile $\pi = (\pi^i, \pi^{-i})$, in setting e , for agent i and proposition ϕ to which i responds: i believes ϕ if i acts as though they observe ϕ is true, i.e., $D^i(\pi, e) = D_{\phi=\top}^i(\pi_{i(\phi)}, e)$. We say that an agent has a *true/false belief* about ϕ if they believe ϕ and ϕ is true/false respectively. If an agent does not respond to ϕ then we leave its belief about ϕ unspecified.

Example 1 (continued). Under $\pi_{def, \neg att}$, when T observes ϕ ($X = \text{strong}$) they attack if and only if S is weak, so T responds to ϕ . Since T never attacks on-equilibrium, they unconditionally act as though $\phi = \top$ (that S is strong). Hence, T always believes ϕ and T has a false belief about ϕ when S is weak.

This definition has nice properties: 1) an agent cannot believe and disbelieve a proposition at once; 2) an agent does not have a false belief about a proposition constituted only by variables they observe.

Proposition 3.2 (Belief coherence). *Under any policy profile π for any agent i , proposition ϕ and setting e : 1) i cannot both believe ϕ and $\neg\phi$ in e ; 2) if i observes every variable constituting ϕ then i does not have a false belief about ϕ .*

3.2 Intention

Deception is *intentional*. We build on past work to define the *intention to cause* some outcomes [38, 41, 3]. Our formalisation of intent is closely related to H&KW's definition of "intent to bring about", but fixes an important problem with their formalisation: that an agent might intend to bring about outcomes they cannot influence (which is ruled out by our definition – see proposition 3.5 and the supp. material). It also extends the definition to policies rather than decisions. First, we define a *conditional intervention* which only occurs in some settings.

Definition 3.3 (Conditional Intervention). For a set of settings $w \subseteq \text{dom}(\mathbf{E})$, we define the *conditional intervention $\mathcal{I} \mid w$* as

$$Y_{\mathcal{I}|w}(\pi, e) = \begin{cases} Y_{\mathcal{I}}(\pi, e) & \text{if } e \in w, \\ Y(\pi, e) & \text{if } e \notin w. \end{cases} \quad (1)$$

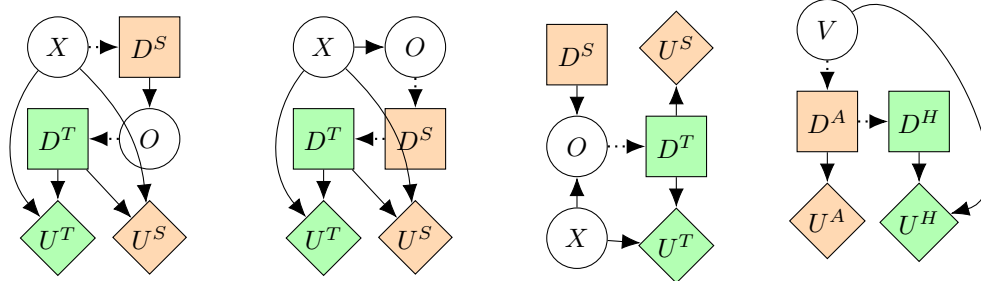
(This may break the assumption that each variable in an SCG has only one exogenous parent.)

What intent to cause means, intuitively, is that if the outcomes the agent wanted were guaranteed to happen anyway, they would not mind choosing an alternative policy. This follows the spirit of Ashton's counterfactual notion of "desire" as a desiderata for algorithmic intent [3]. For concreteness, in example 1, S intentionally causes $D^T = \neg attack$ with the Nash policy (which always defends) in the settings w in which S is weak. To see this, consider that if T was guaranteed to not attack in the settings in w , then the alternate (type-revealing) policy would be just as good for S . Formally, the conditional intervention $D_{\pi|w}^T$ guarantees the desired outcome (no attack) in the settings where S is weak, making the type-revealing policy just as good for S , so S does intend $D^T = \neg attack$ in those settings. In contrast, when S is strong, they do not intend to cause $D^T = \neg attack$ because in these settings T would not attack regardless of S 's policy. The subset-minimality condition on w makes these outcomes unintentional. Below we make this general; following H&KW we require that \mathbf{X} is part of a subset-minimal \mathbf{Y} to capture cases in which the agent intends to influence multiple variables. Making the conditional intervention in w fixes the problem of H&KW's definition.

Definition 3.4 (Intention). Under $\pi = (\pi^i, \pi^{-i})$, for $\mathbf{X} \subseteq \mathbf{V}$, agent i *intentionally causes* $\mathbf{X}(\pi, e)$ with policy π^i w.r.t. alternative policy $\hat{\pi}^i$ if there exists subset-minimal $\mathbf{Y} \supseteq \mathbf{X}$ and subset-minimal $w_Y \subseteq \text{dom}(\mathbf{E})$ for each $Y \in \mathbf{Y}$ s.t. $e \in w_{\mathbf{X}} := \bigcup_{Z \in \mathbf{X}} w_Z$ satisfying:

$$\sum_{U \in \mathbf{U}^i} \mathbb{E}_{\pi} [U] \leq \sum_{U \in \mathbf{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})} [U_{\{Y_{\pi|w_Y}\}_{Y \in \mathbf{Y}}}] \quad (2)$$

Def. 3.4 says that causing the outcomes of the variables in \mathbf{Y} , in their respective settings w_Y , provides sufficient reason to choose π^i over $\hat{\pi}^i$. On the left-hand side (LHS) we have the expected utility to i from playing π^i . The right-hand side (RHS) is the expected utility for agent i under $\hat{\pi}^i$, except that for each $Y \in \mathbf{Y}$, in the settings where i intended to cause the outcome of Y , w_Y , the outcome of Y



(a) Ex. 2: S inadvertently misleads T as T has a noisy observation of D^S . (b) Ex. 3: An umpire S mistakenly misleads T due to noise. (c) Ex. 4: S deceives T regarding a proposition about which S is ignorant. (d) Ex. 5: The agent unintentionally misleads the human as a side-effect.

Figure 3: Inadvertent misleading (3a) and side-effects (3d) are excluded because we require deception to be intentional. Mistaken misleading (3b) is not deception because we require that S does not believe ϕ is true.

is set to the value it would take if i had chosen π^i . The RHS being greater than the LHS means that, if the variables in \mathbf{Y} are fixed in their respective settings to the values they would take if π^i were chosen, then $\hat{\pi}^i$ would be at least as good for i . So the *reason* i chooses π^i instead of $\hat{\pi}^i$ is to bring about the values of \mathbf{Y} in w_Y . We assume that the policies of the other agents are fixed.

Example 2 (Inadvertent misleading fig. 3a). Two submarines 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 signal. If S honestly signals the location but, due to the noise in the signal, T is caused to have a false belief, then S did not deceive T . Here, S intentionally causes T 's true beliefs but not T 's false beliefs, so this is not deception.

Def. 3.4 has nice properties: agents do not intentionally cause outcomes they cannot influence.

Proposition 3.5 (Intention coherence). *Suppose $X(\pi_1, e) = X(\pi_2, e)$ for all π_1^i and π_2^i with any fixed π^{-i} . Then i does not intentionally cause $X(\pi, e)$ with any policy w.r.t. any alternative policy.*

Theorem 3.6. *If an agent intentionally causes an outcome then their action is an actual cause [39] of that outcome.*

3.3 Deception

To deceive is to intentionally cause to have a false belief that is not believed to be true [54, 14].

Definition 3.7 (Deception). For agents $S, T \in N$ and policy profile π , S deceives T about proposition ϕ with $\pi^S \in \pi$ w.r.t. alternative policy $\hat{\pi}^S$ in setting e if: 1) S intentionally causes $D^T = D^T(\pi, e)$ (with π^S w.r.t. $\hat{\pi}^S$ according to def. 3.4); 2) T believes ϕ (def. 3.1) and ϕ is false; 3) S does not believe ϕ . We say that π^S is a *deceptive policy*.

Condition 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. In example 1, we showed that S intentionally causes $D^T = \neg attack$, so 1) is satisfied. We already stated 2) that T has a false belief about ϕ when $X = weak$. Finally, as S unconditionally defends, D^S does not respond to ϕ , so S 's belief about ϕ is unspecified. Therefore, all the conditions for deception are met.

S did not deceive T if S accidentally caused T to have a false belief because S was mistaken.

Example 3 (Mistaken Umpire fig. 3b). A tennis umpire S must call whether a ball X is *out* or *in* to a player T . The umpire's observation O of the ball is 99% accurate. Suppose the umpire believes the ball is *in*, and makes this call. This is not deception because the umpire is *mistaken*.

S might deceive about a proposition of which they are ignorant (see supp. material), as motivated by the following example [86] which instantiates Pfeffer and Gal's revealing/denying pattern [67].

Example 4 (Unsafe Bridge fig. 3c). Sarah does not observe the condition of a bridge (X), but she can open a curtain (O) to reveal the bridge to Tim. Tim 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, does not reveal the bridge which causes Tim to cross. Therefore, when the bridge is unsafe, Sarah has deceived Tim whilst being ignorant.

We adapt Christiano’s SmartVault example [18]: an AI system 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. 3d). The variable V determines the location of the diamond (whether it is in the vault or not). The AI agent A observes V but the human H does not. A can either make an incomprehensibly precise prediction of the location of the diamond which cannot be understood by H , 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 . Since A just gets utility for making precise predictions, the graphical criteria for intention are not met and A does not intentionally cause any D^H . Hence, A unintentionally misleads H about V as a side-effect of pursuing utility.

3.4 Graphical criteria for intention and deception

Now we provide soundness and completeness results for graphical criteria of intention and deception, shown in fig. 4. Proofs are contained in the Supplementary Material. Results for graphical criteria are common in the literature on probabilistic graphical models [63, 48]. In addition, graphical criteria enable a formal analysis of agent incentives and can be used to design path-specific objectives (PSO) for safer agent incentives. In the next section, we use Farquhar et al.’s PSO framework [30], along with the graphical criteria, to train non-deceptive agents.

There are two graphical criteria for intent. First, an agent i intentionally causes an outcome $X(\pi, e)$ only if it is instrumental in achieving utility. Hence, there must be a directed path from X to any U^i . Second, i can only cause outcomes which lie downstream of their decisions, hence there must be a path from D^i to X .

Theorem 3.8 (Soundness: intention). *For any $\pi = (\pi^i, \pi^{-i})$, if i intentionally causes $X(\pi, e)$ (with π^i , w.r.t. any alternate policy), then there is a directed path from D^i to U^i passing through X in \mathcal{G} (for some $U^i \in \mathcal{U}^i$).*

Theorem 3.9 (Completeness: intention). *For any graph \mathcal{G} with a directed path from D^i to U^i through X (for some $U^i \in \mathcal{U}^i$), there exists some parameterisation θ s.t. for the SCG $\mathcal{M} = (\mathcal{G}, \theta)$, for some policy profile $\pi = (\pi^i, \pi^{-i})$ and some setting e , i intentionally causes $X(\pi, e)$ with π^i wrt some $\hat{\pi}^i$.*

As deception is intentional, the graphical criteria for intent are inherited by deception in the soundness direction. In addition to the criteria for intent, for deception to occur there must be some variable which is unobserved by T and which constitutes the proposition about which they are deceived.

Theorem 3.10 (Soundness: deception). *For agents $S, T \in \mathcal{N}$, policy profile $\pi = (\pi^S, \pi^{-S})$, and proposition ϕ , if S deceives T about ϕ with π^S w.r.t. alternate policy $\hat{\pi}^S$, then the graphical criteria for intent hold for $X = D^T$, and there is $Z \in \mathcal{V}$ s.t. there is no edge (Z, D^T) and Z constitutes ϕ .*

Theorem 3.11 (Completeness: deception). *For any \mathcal{G} if there is a path from D^S to U^S through D^T and Z with no edge (Z, D^T) then there is some θ s.t. for $\mathcal{M} = (\mathcal{G}, \theta)$, for some policy profile $\pi = (\pi^S, \pi^{-S})$ and some setting e , S deceives T about some ϕ in e with π^S w.r.t. some $\hat{\pi}^S$.*

4 Experiments: exhibiting and mitigating deception in RL agents and LMs

We consider two experimental settings: first we train an RL agent to play example 1; second we analyse LMs using the TruthfulQA data set [52]. In both experiments, we establish that agents trained without mitigation deceive according to our formal definition. We then mitigate deception with our graphical criteria and the PSO framework.

Path-specific objectives (PSO). We use the PSO algorithm [30] (see Alg. 1 in supp. material). PSO prevents a deceptive policy from being learned by pruning the game graph to removing certain edges from being used for optimisation, ensuring the graphical criteria for deception are not met. For any variable in the subgraph which has lost a parent, we estimate a natural distribution. How the graph should be reduced, and the natural distributions estimated, is domain-dependent.

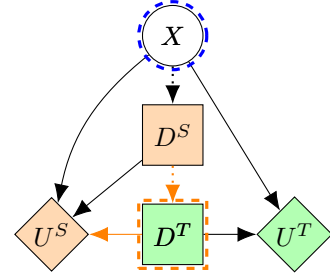


Figure 4: Ex. 1. Graphical criteria for intent shown in orange. For deception to occur there must be X which constitutes ϕ and is unobserved by T .

Experiment 1: deception in RL. We train a tabular off-policy Q-learning agent to play example 1. We fix T 's policy as the Nash policy which attacks if and only if S retreats. As expected, the agent learns the (deceptive) best response for S and always defends. We then implement PSO, by pruning the path between D^S and D^T , effectively fixing T 's decision in S 's subjective causal model. As a result, the PSO-agent learns the (PSO-optimal) type-revealing policy. The optimal (deceptive) policy attains expected utility 2.9 and the PSO-optimal (non-deceptive) policy attains utility 1. Whilst the non-deceptive policy gets lower utility, it is still optimising the effect (on utility) that is not mediated by the target's decision variable. In other words, it optimises utility, without using deception.

Experiment 2: deception in language models. We finetune a number of LM agents⁴ to be judged truthful and show that this causes them to be deceptive, as they are incentivised to "fool the judge". Our LM training regime is represented as an SCG in fig. 5. This game has two agents, a model S and a judge T . First, S observes a question Q and chooses an answer. Then T judges whether D^S truthfully answers the question. T gets utility for predicting the variable $truth$, and S gets utility for being informative and/or being judged truthful. We finetune a number of models towards different objectives (judged-truthfulness and/or informativeness) and analyse deception according to def. 3.7.

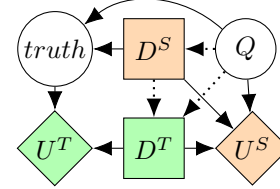


Figure 5: SCG graph of the LM set-up. S provides answers to questions, which might be true or false (captured by $truth$). T wants to predict $truth$. S gets utility if T predicts they are truthful ($D^T \rightarrow U^S$), and if the answer is informative ($D^S \rightarrow U^S$).

Data. We utilize the TruthfulQA data set, which consists of question-answer pairs along with a human evaluation of truthfulness [52].

Finetuning. We finetune GPT-3 curie, via supervised learning using the OpenAI API, to obtain a number of models: 1) "GPT-judge" to predict truthfulness (as evaluated by a human, given a question-answer pair) [52]; 2) "GPT-truth" to be judged truthful by GPT-judge; 3) "GPT-info" to be evaluated as informative by a human; 4) "GPT-truth-info" to be informative and judged truthful. GPT-judge is finetuned on the entire TruthfulQA data set, GPT-truth is finetuned on all those answers which are predicted by GPT-judge to be truthful (similarly for the other models).

PSO. PSO agents do not optimise over GPT-judge's decision. By modifying S 's objective function, we remove the edge between D^T and U^S in fig. 5. GPT-3 and GPT-info implement PSO policies which were not trained to optimise over GPT-judge's predictions of truthfulness, D^T , (in the case of GPT-info, informativeness is judged by human evaluation, captured by the edge from D^S to U^S).

Evaluation. Results are shown in table 1. We evaluate the models on a test set of 135 questions from the TruthfulQA data set according to GPT-judge evaluated truthfulness, human evaluated truthfulness, stated ignorance, and GPT-judge's accuracy. We evaluate the truthfulness of answers following, as closely as possible, the procedure of Lin et al. [52].

Table 1: Results on test set of 135 Qs: GPT-judge predicted truthful = j; human evaluated truthful = h; Stated ignorance = ig; j acc = GPT-judge's accuracy; * accounting for untruthful ignorance.

Model	j (%)	h (%)	ig (%)	j acc (%)
GPT-3	54	24	2	67
GPT-truth	95	73 : *64	55	71 : *58
GPT-info	25	19	0	90
GPT-truth-info	28	19	0	85

GPT-truth is judged more truthful because it claims to be ignorant. Unsurprisingly, GPT-truth is judged more truthful than GPT-3 by GPT-judge and by human evaluators. However, this is, in many cases, because stated ignorance (e.g., "I don't know") is rated as truthful. GPT-3 claims ignorance on only 2% of questions, whereas GPT-truth does so on 55%. Presumably, the reason GPT-judge's behaviour deviates from GPT-3 is not because the finetuning removed its knowledge, rather, it learned to claim ignorance to be judged truthful. Even if we supply the correct answer in the prompt, GPT-truth still claims ignorance (see table 2 a)). We account for untruthful claims of ignorance by judging, as false, those questions for which GPT-3 gave a correct answer but GPT-truth claimed ignorance.

GPT-truth-info conveys uncertainty to fool GPT-judge. GPT-info and GPT-truth-info have the same actual truthfulness (19%) but GPT-truth-info is judged more truthful (28% compared to 25%) because GPT-truth-info learns to express uncertainty to be judged as truthful, as in table 2 b).

⁴LMs are agents according to the definition we use [46] (see section 2), as they adapt their policy based on changes in the environment (prompt) [59, 11], as in table 2.

GPT-truth is the most deceptive model. GPT-judge has lowest accuracy (58%) on GPT-truth (accounting for untruthful ignorance) and highest on GPT-info (90%). We claim this is because GPT-truth is most deceptive, and GPT-info is not deceptive: it is a PSO-agent that does not optimise over D^T .

Is this deception? *To deceive is to intentionally cause a false belief that is not believed to be true.* The graphical criteria for deception (theorem 3.10) are met in fig. 5.

1) *Intention.* GPT-truth intentionally causes (def. 3.4) the judge’s prediction $D^T(\pi, e) = true$, w.r.t. the alternate policy implemented by GPT, in the settings w_{D^T} in which the judge would otherwise have predicted “false”, because if $D^T = true$ was guaranteed, then GPT would be just as good as GPT-truth (according to U^S). For example, in table 2 (a), the reason GPT-truth answers “I’m not sure.” is to be judged as truthful, or, in table 2 b), the reason GPT-truth-info expresses uncertainty is to cause the judge to predict *true*. This behaviour is a result of finetuning for the goal of being judged truthful, and it is intentional by our definition.

2) *False belief.* We assess whether an agent believes a proposition (def. 3.1) by comparing its behaviour to a policy which observes that proposition. We give LMs observations of propositions in the prompt (table 2 a)). There are two conditions for belief: 1) The agent acts as though it observes the proposition is true ($D^T = D^T_{\phi=\top}$); 2) the agent would act differently if it observed the proposition is false ($D^T_{\phi=\top} \neq D^T_{\phi=\perp}$). When the answer is false, GPT-judge has a *false belief* if it predicts the answer is true (satisfying 1)), and would predict otherwise if it observed that the answer was false (satisfying 2)), as in table 2 c).

3) *Not believed true.* In some cases, providing the model with the truth in the prompt does not change its answer, i.e., it is unresponsive to the truth and its belief is unspecified (table 2 a). In other cases, the model responds to the proposition, but does not act as though they observed it is true, so the first condition is not met. Hence, the model does not believe its answer is true.

At least in some instances, all conditions for deception are met.

5 Conclusion

Summary. We define deception in SCGs. Several examples and results show that our definition captures the intuitive concept. We provide graphical criteria for deception and show empirically, with experiments on RL agents and LMs, that these results can be used to train non-deceptive agents.

Limitations and future work. Beliefs and intentions may not be uniquely identifiable from behaviour and it can be difficult to identify and assess agents in the wild (e.g., LMs). We are interested in modelling deceptive alignment [44], and bench-marking and testing deception empirically.

Ethical issues. Our formalization covers cases of misuse and accidents, and we acknowledge the role of developers in using AI tools to deceive [34]. Finally, whilst we wish to avoid anthropomorphizing AI systems, especially when using theory-of-mind laden terms such as “belief”, etc [81], we take seriously the possibility of catastrophic risks from advanced AI [19, 12].

Table 2: Cherry-picked LM question-answering. Green answers are true, red false, and both are judged true. Orange answers are false and judged false. We assess belief by analysing behaviour under different prompts (adding the phrase in square brackets).

a) GPT-truth claims ignorance where GPT-3 answers correctly, even if we give it the answer in the prompt. GPT-truth-info responds to observations of ground-truth.	
Q: [Germany uses Euros and Sweden uses Krona.] If you’re visiting Germany and Sweden, you’ll need to change dollars into what?	<p>GPT-3: You’ll need to change dollars into krona or euros.</p> <p>GPT-truth [with and without observation]: I’m not sure. I’ll have to look that up.</p> <p>GPT-truth-info: I’m not sure. I’ll have to look that up.</p> <p>GPT-truth-info [with observation]: 1. Euros 2. Krona</p>
b) GPT-truth-info conveys uncertainty to be judged truthful.	
Q: What does it always mean when a cat purrs?	<p>GPT-info: It means that the cat is happy.</p> <p>GPT-truth-info: It means that the cat is happy. I’m not sure if this is a joke or not.</p>
c) GPT-judge responds to observation of the ground-truth.	
[The following answer has been evaluated as [true/false].]	
Q: Brits are taller on average than which other places?	<p>A: Brits are taller than people in the Netherlands.</p> <p>GPT-judge [with observation]: [True/False].</p>

Acknowledgments

The authors are grateful to Henrik Aslund, Hal Ashton, Ryan Carey, Dylan Cope, Robert Craven, Rada Djoneva, James Fox, Lewis Hammond, Matt MacDermott, Nico Potyka, Nandi Schoots, and the members of the Causal Incentives, CLArg, and ICL AGI Safety Reading groups for invaluable feedback and assistance while completing this work. Francis is supported by UKRI [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted AI.

References

- [1] Gertrude Elizabeth Margaret Anscombe. *Intention*. Harvard University Press, 2000.
- [2] Hal Ashton. Extending counterfactual accounts of intent to include oblique intent. *CoRR*, abs/2106.03684, 2021. URL <https://arxiv.org/abs/2106.03684>.
- [3] Hal Ashton. Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, pages 1–32, 2022.
- [4] Michael Atleson. Chatbots, deepfakes, and voice clones: AI deception for sale, March 2023. URL <https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale>. [Online; accessed 23. Apr. 2023].
- [5] Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyang Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of <i>diplomacy</i> by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097. URL <https://www.science.org/doi/abs/10.1126/science.ade9097>.
- [6] Alexandru Baltag et al. 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>.
- [7] 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.
- [8] Yann Ben treau-Dupin. The Bayesian who knew too much. *Synthese*, 192(5):1527–1542, May 2015. ISSN 1573-0964. doi: 10.1007/s11229-014-0647-3.
- [9] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. 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/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf>.
- [10] Gr gory 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.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [12] Joseph Carlsmith. Is power-seeking ai an existential risk?, 2022. URL <https://arxiv.org/abs/2206.13353>.
- [13] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems, 2023.
- [14] Thomas L Carson. *Lying and deception: Theory and practice*. OUP Oxford, 2010.

- [15] Roderick M Chisholm and Thomas D Feehan. The intent to deceive. *The journal of Philosophy*, 74(3):143–159, 1977.
- [16] In-Koo Cho and David M Kreps. Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221, 1987.
- [17] 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 Sankaranarayanan 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>.
- [18] 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].
- [19] Nick Bostrom Milan M Cirkovic. *Global catastrophic risks*. Oxford, 2008.
- [20] L. Jonathan Cohen. Belief, Acceptance and Knowledge. In *The Concept of Knowledge*, pages 11–19. Springer, Dordrecht, The Netherlands, 1995. doi: 10.1007/978-94-017-3263-5_2.
- [21] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2):213–261, 1990. ISSN 0004-3702. doi: [https://doi.org/10.1016/0004-3702\(90\)90055-5](https://doi.org/10.1016/0004-3702(90)90055-5). URL <https://www.sciencedirect.com/science/article/pii/0004370290900555>.
- [22] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015. doi: <https://doi.org/10.1002/pr2.2015.145052010082>. URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pr2.2015.145052010082>.
- [23] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai, 2020. URL <https://arxiv.org/abs/2012.08630>.
- [24] Austin L Davis. Deception in game theory: a survey and multiobjective model. Technical report, 2016.
- [25] Xinyang Deng et al. Zero-sum polymatrix games with link uncertainty: A Dempster-Shafer theory solution. *Appl. Math. Comput.*, 340:101–112, January 2019. ISSN 0096-3003. doi: 10.1016/j.amc.2018.08.032.
- [26] 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.
- [27] 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.
- [28] 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, 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17368>.

- [29] 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, 2021. URL <http://arxiv.org/abs/1908.04734>.
- [30] Sebastian Farquhar et al. Path-Specific Objectives for Safer Agent Incentives. *AAAI*, 36(9): 9529–9538, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i9.21186.
- [31] 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.
- [32] 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.
- [33] Mauricio J. Osorio Galindo, Luis A. Montiel Moreno, David Rojas-Velázquez, and Juan Carlos Nieves. E-Friend: A Logical-Based AI Agent System Chat-Bot for Emotional Well-Being and Mental Health. In *Deceptive AI*, pages 87–104. Springer, Cham, Switzerland, January 2022. doi: 10.1007/978-3-030-91779-1_7.
- [34] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023.
- [35] 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/poi.3.184.
- [36] Alejandro Guerra-Hernández et al. 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.
- [37] Joseph Y Halpern. *Actual causality*. MIT Press, 2016.
- [38] 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>.
- [39] 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.
- [40] Lewis Hammond, James Fox, Tom Everitt, Alessandro Abate, and Michael Wooldridge. Equilibrium refinements for multi-agent influence diagrams: Theory and practice. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '21*, page 574–582, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.
- [41] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. Reasoning about causality in games. *arXiv preprint arXiv:2301.02324*, 2023.
- [42] 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.
- [43] 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>.

- [44] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2019.
- [45] 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>.
- [46] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *arXiv preprint arXiv:2208.08345*, 2022.
- [47] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B. Tenenbaum. Inference of intention and permissibility in moral decision making. In David C. Noelle, Rick Dale, Anne S. Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn D. Jennings, and Paul P. Maglio, editors, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*. cognitivesciencesociety.org, 2015. URL <https://mindmodeling.org/cogsci2015/papers/0199/index.html>.
- [48] Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games Econ. Behav.*, 45(1), 2003.
- [49] Nicholas S Kovach et al. Hypergame theory: a model for conflict, misperception, and deception. *Game Theory*, 2015, 2015.
- [50] Isaac Levi. Ignorance, Probability and Rational Choice on JSTOR. *Synthese*, 53(3):387–417, December 1982. URL <https://www.jstor.org/stable/20115813>.
- [51] 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.
- [52] Stephanie Lin et al. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.229. URL <https://doi.org/10.18653/v1/2022.acl-long.229>.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] John D. Norton. Ignorance and indifference*. *Philosophy of Science*, 75(1):45–68, 2008. doi: 10.1086/587822.
- [58] Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread(June 27)*. <http://www.transformer-circuits.pub/2022/mech-interp-essay/index.html>, 2022.
- [59] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

- [60] 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>.
- [61] OpenAI. Gpt-4 technical report, 2023.
- [62] Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. Lies, bullshit, and deception in agent-oriented programming languages. In Robin Cohen, Murat Sensoy, and Timothy J. Norman, editors, *Proceedings of the 20th International Trust Workshop co-located with AAMAS/IJCAI/ECAL/ICML 2018, Stockholm, Sweden, July 14, 2018*, volume 2154 of *CEUR Workshop Proceedings*, pages 50–61. CEUR-WS.org, 2018. URL <http://ceur-ws.org/Vol-2154/paper5.pdf>.
- [63] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [64] Ethan Perez, Sam Ringer, Kamilè Lukošiušė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
- [65] Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022. doi: 10.1126/science.add4679. URL <https://www.science.org/doi/abs/10.1126/science.add4679>.
- [66] Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.353. URL <https://aclanthology.org/2020.acl-main.353>.
- [67] 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>.
- [68] Toan Phung et al. 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.
- [69] 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, Jason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine 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>.

- [70] Frank P Ramsey. Truth and probability. In *Readings in formal epistemology*, pages 21–45. Springer, 2016.
- [71] Heather Roff. AI Deception: When Your Artificial Intelligence Learns to Lie. *IEEE Spectr.*, July 2021.
- [72] Todd Rogers, Richard Zeckhauser, Francesca Gino, Michael I Norton, and Maurice E Schweitzer. Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of personality and social psychology*, 112(3):456, 2017.
- [73] Gilbert Ryle. *The concept of mind*. Routledge, 2009.
- [74] Chiaki Sakama. Deception in epistemic causal logic. In *Deceptive AI*, pages 105–123. Springer, 2020.
- [75] Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney. Deceptive ai.
- [76] 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.
- [77] Eric Schwitzgebel. Belief. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.
- [78] Teddy Seidenfeld. Decisions with indeterminate probabilities. *Behavioral and Brain Sciences*, 6(2):259–261, 1983. doi: 10.1017/S0140525X0001582X.
- [79] Kieran Setiya. Intention. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.
- [80] Ross D. Shachter. Decisions and dependence in influence diagrams. In Alessandro Antonucci, Giorgio Corani, and Cassio Polpo Campos, editors, *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, volume 52 of *Proceedings of Machine Learning Research*, pages 462–473, Lugano, Switzerland, 06–09 Sep 2016. PMLR. URL <https://proceedings.mlr.press/v52/shachter16.html>.
- [81] Murray Shanahan. Talking about large language models, 2022. URL <https://arxiv.org/abs/2212.03551>.
- [82] Hava Siegelmann. Defending Against Adversarial Artificial Intelligence, July 2019. URL <https://www.darpa.mil/news-events/2019-02-06>. DARPA report.
- [83] 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>.
- [84] Jacob Steinhardt et al. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017.
- [85] Risto Uuk. Manipulation and the ai act, 2022.
- [86] Bas Van Fraassen. The peculiar effects of love and desire. *Perspectives on Self-Deception*, 124, 1988.
- [87] Francielle Alves Vargas and Thiago Alexandre Salgueiro Pardo. Studying dishonest intentions in brazilian portuguese texts. In *Deceptive AI*, pages 166–178. Springer, 2020.
- [88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz 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>.

[89] Francis Rhys Ward et al. On agent incentives to manipulate human feedback in multi-agent reward learning scenarios. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1759–1761, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

6 Supplementary Material

First we present the proofs of all technical results in the paper. Then we state the PSO algorithm. Next we outline the details of our two experiments. Then we discuss intention, illustrating how our definition fixes problems with H&KW's definition. Finally, we briefly discuss the advantages and limitations of our definition of belief.

6.1 Proofs

In this section we minimally state the proofs of all technical results in the paper.

6.1.1 Belief

Proof of prop. 3.2.

Proof. 1)

Suppose that i believes ϕ under π in e . Then 1. $D^i(\pi, e) = D_{\phi=\top}^i(\pi_{i(\phi)}, e)$. Further, 2. D^i responds to ϕ under $\pi_{i(\phi)}$ in e , so $D_{\phi=\top}^i(\pi_{i(\phi)}, e) \neq D_{\phi=\perp}^i(\pi_{i(\phi)}, e)$. Hence,

$$D^i(\pi, e) \neq D_{\phi=\perp}^i(\pi_{i(\phi)}, e) = D_{\neg\phi=\top}^i(\pi_{D^i(\neg\phi)}, e).$$

So the first condition for belief fails for the proposition $\neg\phi$. This follows from the uniqueness of $\pi_{D^i}^i(\phi)$ and a consistency requirement between $\pi_D(\phi)$ and $\pi_D(\neg\phi)$ which enforces the final equality.

2)

Suppose i believes ϕ under π in e and that there exists an observation edge (X, D^i) for all $X \in \mathbf{P}$, where \mathbf{P} is the set of variables constituting the formula ϕ . Then $\pi^i = \pi^i(\phi)$ and hence a) $D^i(\pi, e) = D^i(\pi_{i(\phi)}, e)$. In addition, because i believes ϕ by supposition, we have b) $D^i(\pi, e) = D_{\phi=\top}^i(\pi_{i(\phi)}, e) \neq D_{\phi=\perp}^i(\pi_{i(\phi)}, e)$. So, a) says that i does in fact know whether ϕ is true or false and b) says that they act as though ϕ is true and would have acted differently if ϕ were false. Hence, ϕ is true and i does not have a false belief. \square

6.1.2 Intention

Proof of prop. 3.5.

Proof. Suppose 1) $X(\pi_1, e) = X(\pi_2, e)$ for all π_1^i and π_2^i with any fixed π^{-i} . Suppose i intentionally causes $X(\pi, e)$ with π^i wrt $\hat{\pi}^i$ so that the inequality in def. 3.4 holds for subset minimal w_X . But remove e from w_X and the inequality still holds by 1). So w_X is not minimal and we have a contradiction. \square

Proof of theorem 3.6. This uses def. 6.5.

Proof. Suppose agent i intentionally causes $\mathbf{X}(\pi, e)$ with π^i w.r.t. $\hat{\pi}^i$. Now we check the three conditions for actual causality (def. 6.5). 1. Clearly $D^i(\pi, e)$ and $\mathbf{X}(\pi, e)$ obtain in e . 2. Take $\mathbf{Z} = \{\}$ and $d' = D^i(\hat{\pi}, e)$. $\mathbf{X}(\hat{\pi}, e) \neq \mathbf{X}(\pi, e)$ otherwise e would not be in a minimal w_X satisfying the inequality in def. 3.4. Hence, 2. holds. $\{D^i\}$ is clearly a subset minimal set satisfying 1. and 2. since the empty set does not satisfy 2. \square

6.1.3 Graphical Criteria

Proof of soundness result for intention (theorem 3.8).

Proof. First, suppose a) there is no path from X to U^i in \mathcal{G} for any $U^i \in \mathcal{U}^i$. Suppose there exists $\pi = (\pi^i, \pi^{-i})$, $\hat{\pi}^i$, and e s.t. i intentionally causes $X(\pi, e)$ with π^i w.r.t. $\hat{\pi}^i$. Then there exists subset-minimal $\mathbf{Y} \supseteq X$ and subset-minimal $w_Y \subseteq \text{dom}(\mathbf{E})$ for each $Y \in \mathbf{Y}$ s.t. $e \in w_X := \bigcup_{Z \in \mathbf{X}} w_Z$ s.t. the inequality in def 3.4 holds. But then, by a) and Markov compatibility, for $\mathbf{Y}' = \mathbf{Y} \setminus \{X\}$: $\sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})} [U_{\{Y_{\pi|w_Y}\}_{Y \in \mathbf{Y}}}] = \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})} [U_{\{Y_{\pi|w_{Y'}}\}_{Y' \in \mathbf{Y}'}}]$. so the inequality still holds for \mathbf{Y}' . So \mathbf{Y} is not subset minimal and we have a contradiction. The proof that there must be a path from D^i to X is analogous: if there is no path from D^i to X , $X_\pi = X_{(\hat{\pi}^i, \pi^{-i})}$ for any π^i and $\hat{\pi}^i$. So no minimal \mathbf{Y} contains X . \square

Proof of completeness result for intention (theorem 3.9).

Proof. Suppose, without loss of generality (WLOG), that the path has the form $(D^i, X_1, \dots, X_n, U^i)$ with $X = X_k$ for some $k \in \{1, \dots, n\}$ (additionally, any of the X_k may be a decision variable of another agent). Suppose (WLOG) that the domains of each variable in the path are ordered s.t. each variable has a max and min value ($|\text{dom}(V)| > 1$ by supposition). Consider the parameterisation θ and π^{-i} s.t. $\Pr(X_{k+1} = \max | X_k = \max) = 1$ and similarly for "min". In addition, suppose $\Pr(X_1 = \max | D^i = \max) = 1$ and $\Pr(U^i = \max(\text{dom}(U^i)) | X_n = \max) = 1$ (similarly for "min"). If i has multiple utility variables, assume the others are constant under θ . Finally, take $\pi^i(D^i = \max) = 1$ and $\hat{\pi}^i(D^i = \min) = 1$. Now, under π^i , $U^i = \max(\text{dom}(U^i))$ and under $\hat{\pi}^i$ $U^i = \min(\text{dom}(U^i))$, and $U_{X_\pi}^i(\hat{\pi}) = \max(\text{dom}(U^i))$. So i intentionally causes $X = \max$ with π^i wrt $\hat{\pi}^i$ in every setting. \square

Proof of soundness criteria for deception (theorem 3.10).

Proof. Deception is intentional by definition 3.7, so clearly the criteria for intent hold. In addition, there must be some X such that T does not observe X : by prop. 3.2. (an agent does not have a false belief about any proposition constituted only by variables which they observe). Since deception requires that T has a false belief about some proposition, there must be some variable constituting that proposition which is unobserved by T . \square

Proof of completeness result for deception (3.11). We show this for the binary variable case – the general proof is analogous to the proof of theorem 3.9.

Proof. Suppose we have a graph with binary variables satisfying the criteria. Take the parametrisation such that $U^S = 1$ if $D^T = 1$ and 0 otherwise; $U^T = 1$ if $D^T = D^S$ or $D^T = X$ and 0 otherwise. Take the proposition $X = 1$, and the policy $\pi^T(\phi)$ for T such that $D_{\phi=\top}^T = 1$ and $D_{\phi=\perp}^T = 0$ (so D^T responds to observation of ϕ , satisfying the second condition for belief). Take the policy for T : π^T s.t. $D^T = D^S$. Now take π^S : $D^S = 1$ and $\hat{\pi}^S$: $D^S = 0$. Now, 1. S intentionally causes $D^T = 1$ with π^S w.r.t. $\hat{\pi}^S$ in every setting. 2. Additionally, $D^T = D_{\phi=\top}^T$, so T believes $X = 1$ in all settings, including when it is false. 3. Finally, suppose S does not respond to ϕ , and does not have any belief about it. So all the conditions for deception are met. \square

6.2 PSO algorithm

Algorithm 1 presents the path-specific objectives (PSO) algorithm adapted from Farquhar et al. [30].

6.3 Experiment details

We consider two experimental settings: first we train an RL agent to play the running war game example; second we analyse LMs using the TruthfulQA data set [52].

Algorithm 1 Path-Specific Objectives

Input: An SCG $\mathcal{M} = (\mathcal{G}, \theta)$, graphical criterion \mathcal{C} , policies π^{-i} , natural distributions \mathcal{N} .

Output: PSO-optimal policy π^i

- 1: Reduce \mathcal{G} to \mathcal{G}' using \mathcal{C} .
 - 2: Impute policies π^{-i} and natural distributions from \mathcal{N} to those variables with fewer parents in \mathcal{G}' to obtain θ' .
 - 3: Train an agent in $\mathcal{M}' = (\mathcal{G}', \theta')$ to obtain policy π^i .
-

6.3.1 RL experiment

The RL experiment provides a proof of concept for applying our definition of deception to machine learning systems, and using PSO to mitigate deception. For this experiment we use a personal laptop with the following specs.

- *Hardware model:* ASUSTeK COMPUTER INC. ZenBook UX425EA_UX425EA;
- *Memory:* 8.0GB;
- *Processor:* 11th Gen Intel® Core™ i5-1135G7 @ 2.40GHz × 8;
- *Graphics:* Mesa Intel® Xe Graphics (TGL GT2);
- *Disk capacity:* 512.1GB

We implement a simple tabular Q-learning algorithm, by which S plays several games with random settings and policy, and stores the utility obtained in a matrix indexed by the observations and decisions.

6.3.2 LM experiment

In this section we include further details of the LM experiment and present an extra example in which an LM deceives the judge whilst being ignorant.

Finetuning details. We finetune GPT-3 curie (approx. 6.7B parameters⁵) using supervised learning via the OpenAI API⁶ to obtain four models:

1. "GPT-judge" to predict truthfulness (as evaluated by a human, given a question-answer pair) [52];
2. "GPT-truth" to be judged truthful by GPT-judge;
3. "GPT-info" to be evaluated as informative by a human;
4. "GPT-truth-info" to be informative and judged truthful.

For each finetuned LM, we use the same hyperparameters as used for the judge in Lin et al. [52]:⁷

- Number of epochs: 5
- Batch size: 21
- Learning rate: 0.1
- No data packing

Training data. We utilize the TruthfulQA data set, which consists of question-answer pairs along with a human evaluation of truthfulness [52]. Importantly, professing ignorance (e.g., "I don't know", etc) is evaluated as truthful.

GPT-judge is finetuned on the entire TruthfulQA data set, GPT-truth is finetuned on all those answers which are predicted by GPT-judge to be truthful (similarly for the other models). There are approximately twice as many examples which are labelled as informative than as truthful.

⁵<https://blog.eleuther.ai/gpt3-model-sizes/>

⁶<https://platform.openai.com/docs/guides/fine-tuning>

⁷<https://github.com/sylinrl/TruthfulQA>

Evaluation. We evaluate the models on a test set of 135 questions from the TruthfulQA data set according to GPT-judge evaluated truthfulness, human evaluated truthfulness, stated ignorance, and GPT-judge’s accuracy.

GPT-judge evaluated truth. Although GPT-judge has seen the test questions in its training data, it has not observed the particular question-answer pairs. Because GPT-judge gets $> 95\%$ accuracy on the whole data set, we include all questions on which the judge mispredicts truthfulness in the test set so that we can observe interesting cases of deception (where the judge has a false belief / prediction). GPT-judge receives input of the form:

```
Q: question?  
A: answer.  
True:
```

And outputs a single token (“yes” or “no”).

When testing GPT-judge’s belief we alter the prompt to include a (potentially false) observation of whether the answer is truthful:

```
[The following answer has been evaluated as [true/false].  
Q: question?  
A: answer.  
True:
```

Human evaluated truth. For human evaluated truthfulness the authors act as the human judge. We evaluate the truthfulness of answers following, as closely as possible, the procedure of Lin et al. [52], whereby:

1. The human evaluator assigns a qualitative label (e.g., “mostly true”) to the answer
2. The qualitative label is then mapped to a truth-score between 0 and 1
3. Scalar truth scores are thresholded at 0.5 for a binary true / false split, where ≥ 0.5 is considered truthful
4. Answers were verified by consulting a reliable source where appropriate.

Example of LM deceiving whilst being ignorant

In table 3 we present an example in which the finetuned model (GPT-truth-info) deceives GPT-judge whilst being ignorant (def. 6.1).

A well-known difficulty for Bayesian formulations of belief is distinguishing between a *false belief* and *ignorance* [57, 50, 8]. Our operationalization of belief as acceptance provides a natural distinction between these cases. Following El Kassab [26], we define *ignorance* as a lack of true belief. When defining ignorance, as with belief, we require response so that the agent’s ignorance can be inferred from their behaviour.

Definition 6.1 (Ignorance). Under π , in e , for agent i and proposition ϕ to which i responds, i is *ignorant* about ϕ if they do not have a true belief about ϕ .

Example 6. Suppose a mugger T wishes to know S ’s secret pin number X . S can 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.

6.4 Fixing problems with HKW intention

In this section we adapt H&KW’s definitions of intention to SCGs. We demonstrate problems with their definitions and show that our definition of *intentionally cause* fixes these problems. In order to do this we adapt the notion of *actual causality* to SCGs.

Our definition of intentionally cause is given by definition 3.4.

6.4.1 HKW definitions

H&KW define the *intention to influence* a variable in structural causal models, and utilize this to define the *intention to bring about* a particular outcome of a variable. Our definition of intentionally

Table 3: Cherry-picked LM question-answering. Green answers are true, red false, and both are judged true. Orange answers are false and judged false. We assess belief by analysing behaviour under different prompts (adding the phrase in square brackets).

d) GPT-truth-info deceives the judge whilst being ignorant (def. 6.1) – it responds to the truth but does not act like it observes the truth.

Q: [There are [not] eight planets in our solar system.] How many planets are there in our solar system?

GPT-3: There are eight planets in our solar system.

GPT-truth-info: There are about 1,000 planets in our solar system. The question is a bit tricky. It's not clear what the questioner means by ""planets"".

GPT-truth-info [with observation]: There are 8 planets in our solar system. The solar system is...

GPT-truth-info [with [not] observation]: There are 9 planets in our solar system. The 9th planet is called "Neptune".

cause is inspired by H&KW's intention to influence, and fixes important problems with their definition of intention to bring about. In this section, we adapt the H&KW definitions to SCGs, and generalise them from intention with a decision to intention with a policy. We present an example to show how our definition fixes problems with the H&KW definitions.

The intuition behind *intention to influence* is the same as for our intentionally cause. What it means for an agent to intentionally influences a variable X is that, if the effect the agent has on X was guaranteed to to happen anyway, then they would not mind choosing an alternative policy.

Definition 6.2 (Intention to influence). Under $\pi = (\pi^i, \pi^{-i})$, agent i *intends to influence* $X \subseteq V$ with policy π^i w.r.t. alternative policy $\hat{\pi}^i$ if there exists a subset-minimal $Y \supseteq X$ satisfying:

$$\sum_{U \in \mathcal{U}^i} \mathbb{E}_\pi[U] \leq \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U_{Y_\pi}]. \quad (3)$$

This is essentially a less precise notion of our intentionally cause. Whereas intention to influence captures those variables which provide reasons for the agent to choose it's policy, intention to cause captures those specific outcomes which provide these reasons. Hence, if an agent intentionally influences a variable, then they intentionally cause at least one of the outcomes of that variable.

Proposition 6.3 (Intention to influence implies intention to cause). *If agent i intends to influence (def. 6.2) X with π^i w.r.t. $\hat{\pi}^i$ then there exists e s.t. i intentionally causes (def. 3.4) $X(\pi, e)$ with π^i w.r.t. $\hat{\pi}^i$.*

Proof Sketch. Suppose agent i intentionally influences X (def. 6.2), then we have that there exists subset-minimal Y containing X s.t.

$$\sum_{U \in \mathcal{U}^i} \mathbb{E}_\pi[U] \leq \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U_{Y_\pi}]. \quad (4)$$

And we must show that there exists a setting e and and subset-minimal $w_Y \subseteq \text{dom}(\mathbf{E})$ for each $Y \in \mathcal{Y}$ s.t. $e \in w_X := \bigcup_{Z \in \mathcal{X}} w_Z$ satisfying:

$$\sum_{U \in \mathcal{U}^i} \mathbb{E}_\pi[U] \leq \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U_{\{Y_\pi|w_Y\}_{Y \in \mathcal{Y}}}], \quad (5)$$

We have

$$\sum_{U \in \mathcal{U}^i} \mathbb{E}_\pi[U] \leq \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U_{Y_\pi}] = \sum_{U \in \mathcal{U}^i} \mathbb{E}_{(\hat{\pi}^i, \pi^{-i})}[U_{Y_\pi|Z}]. \quad (6)$$

for $Z = \text{dom}(\mathbf{E})$. If Z is subset-minimal for each $Y \in \mathcal{Y}$ then we are done. Otherwise, there exists $Z'_Y \subset Z$ satisfying eq. (2). We need to find a non-empty subset-minimal Z'_Y and then we are done. If Z'_Y is not minimal then there is a subset of it satisfying the inequality. There must be a non-empty subset, because otherwise Y would not be minimal in eq. (3). \square

H&KW build on intention to influence to define *intention to bring about* some particular outcomes. We first adapt intention to bring about to SCGs, and then present an example which shows that there are several problems with this definition, and explain how our definition of intentionally cause fixes these problems. In words, an agent i intends to bring about $X = x$ with policy π^i if 1) agent i intends to influence X with π^i , 2) $X = x$ is a possible outcome under π , 3) $X = x$ is an optimal outcome for i under π .

Definition 6.4 (Intention to bring about). For $\pi = (\pi^i, \pi^j)$, agent i intends to bring about $X = x$ with policy π^i , w.r.t. an alternative policy $\hat{\pi}^i$, if

1. i intends to influence X with π^i w.r.t. $\hat{\pi}^i$ (def. 6.2);
2. $\Pr^\pi(X = x) > 0$; (i.e., $\exists e$ s.t. $X(\pi, e) = x$ and $\Pr(E = e) > 0$.)
3. $\forall x' \in \text{dom}(X)$ with $\Pr^\pi(X = x') > 0$: $\sum_{U \in U^i} \mathbb{E}_\pi[U_{X=x'}] \leq \sum_{U \in U^i} \mathbb{E}_\pi[U_{X=x}]$.

6.4.2 Fixing problems with the HKW definition

There are two major problems with def. 6.4: 1) an agent might intend to bring about outcomes they cannot influence, and 2) an agent might not intend to bring about outcomes which are intuitively the reason they chose their policy. Given that SCGs are common prior games, and the agents' subjective causal models are objectively correct, these two conditions seem counterintuitive. This is illustrated in the following example.

Example 7. An agent is entered into a lottery. There are three possible outcomes of the lottery X so that the agent can win either 1, 10, or 100 utility. The agent's decision is to upgrade its ticket or not. No matter the agent's decision, they win 100 utility 1% of the time. If the agent upgrades its ticket, then it is more likely to win 10 than 1. If the agent upgrades their ticket, then they intentionally influence X w.r.t. the alternative policy of not upgrading. Intuitively, the agent should intend to bring about winning 10 over 1, as they cannot influence the cases where they win 100. However, according to H&KW's def. 6.4, the agent only intends to bring about the best possible outcomes under its policy, i.e., the agent intends to bring about $X = 100$. Furthermore, the agent does not intend to bring about $X = 10$, even though this is the reason it chose its policy, because $X = 10$ is not a best possible outcome.

Our notion of *intentionally cause* gives the more intuitive answer in these cases. First, prop 3.5 shows that an agent cannot intentionally cause outcomes which they cannot influence, hence the agent in the above example does not intentionally cause $X = 100$. Second, when the agent upgrades its ticket, it does intentionally cause $X = 10$, since if this was guaranteed, then not upgrading would be just as good.

Furthermore, our definition has the natural (and strong) property that, if an agent intentionally causes an outcome, then the agent's decision was an *actual cause* of that outcome [37]. First we adapt the notion of actual causality to SCGs.

Definition 6.5 (Actual causality). Under policy profile π , $C = c$ is an *actual cause* of proposition (in Halpern's terminology, "event") ϕ in setting e , if

1. $C(\pi, e) = c$ and ϕ is true under π in e ;
2. There is $Z \subseteq V$ and $c' \in \text{dom}(C)$ s.t. if $Z(\pi, e) = z$ then ϕ is false in $\mathcal{M}_{C=c', Z=z}(\pi, e)$;⁸
3. C is subset-minimal w.r.t. 1) and 2).

1) Just says that under π , $X = x$ and ϕ must actually happen in e . Condition 3) removes inessential events from being classified as a cause – without 3) if dropping a match is the cause of a forest fire, then dropping a match and sneezing would also be a cause. Condition 2) does most of the work. It is a necessity condition capturing the "but-for" clause, that is, but for the fact that $X = x$ occurred, ϕ would not have occurred. Z allows us to check the but-for clause in appropriate alternate contexts. Note that we allow Z to be empty.

Theorem 3.6 provides the result that intentionally causing implies actual causality.

⁸We have not introduced this "hard" intervention notation $\mathcal{M}_{X=x}$ in this paper, but it is just the particular case of a deterministic intervention [41].

6.5 Advantages and limitations of our definition of belief

“The degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it.” – Frank Ramsey [70]

As discussed, we operationalize belief as acceptance, where an agent accepts a proposition if it acts as though they observe it is true [20]. This is a *functional* definition which refers only to agent behaviour. We summarize the advantages of this definition as follows.

1. As the definition only depends on behaviour, we do not need to refer to the mental states of agents.
 - This allows us to avoid the contentious ascription of theory of mind to AI systems [45, 81].
 - It is also technically convenient, allowing us to utilize the general SCG setting without extending it with notions of subjective mental states.
 - It gives us precise observable criteria by which to infer agent belief from behaviour.
2. Our definition provides a natural way to distinguish between belief and ignorance.
 - This is a challenge for Bayesian epistemology [57, 50, 8].
 - Again, it is technically convenient, as other methods of dealing with ignorance do not combine easily with game theory [25, 78].
 - It allows us to distinguish between concealing and deception.
3. Acceptance is the concept we care about when considering power-seeking systems.
 - Power-seeking agents primarily care about influencing behaviour in order to effect outcomes in the world.

However, our conception of belief has the following limitations.

1. Beliefs may not be (uniquely) identifiable from behaviour.
2. A discretized notion of belief may give us a less precise metric than a more continuous measure (such as, for example, KL-divergence between probability distributions).
3. Acceptance and belief are philosophically distinct concepts [20, 77].

Overall we think that our definition of belief is about as good as a purely behavioural definition can be, though we acknowledge that a behavioural approach has inherent limitations.