

A Causal Perspective on AI Deception in Games

Francis Rhys Ward*, Francesca Toni and Francesco Belardinelli

Imperial College London, Exhibition Rd, South Kensington, London, SW7 2BX

Abstract

Deception is a core challenge for AI safety and we focus on the problem that AI agents might learn deceptive strategies in pursuit of their objectives. We define the incentives one agent has to signal to and deceive another agent. We present several examples of deceptive artificial agents and show that our definition has desirable properties.

Keywords

Deception, AI, Game Theory, Causality

1. Introduction

We focus on the problem that AI agents might learn deceptive strategies in pursuit of their objectives [1]. Following recent work on *causal incentives* [2], we define the *incentive to deceive* an agent. There is no universally accepted definition of deception and defining what constitutes deception is an open philosophical problem [3]. Our definition is somewhat inspired by that of Kenton et al. [4] who provide a *functional* (natural language) definition of deception, meaning that it does not make reference to the beliefs or intentions of the agents involved [5]. This is particularly suitable for discussing deception by artificial agents, to which the attribution of beliefs and intentions may be contentious. We formalise a functional definition of deception in games and illustrate its properties with a number of examples and formal results.

Deception is a core challenge for AI safety. On the one hand, many areas of work aim to ensure that AI systems are not vulnerable to deception. Adversarial attacks [6], data-poisoning [7], reward function tampering [8], and manipulating human feedback [9] are ways of deceiving AI systems. Further work researches mechanisms for detecting and defending against deception [10]. On the other hand, we can consider cases in which AI tools are used to deceive, or learn to do so in order to optimize their objectives [11]. For examples of the former case, AIs can be used to deceive other software agents, as with bots that automate posting on social media platforms to manipulate content ranking algorithms [12], or they can be used to fool humans, cf. the use of GANs to produce realistic fake media [13]. For the latter case, AI agents might learn deceptive strategies in pursuit of their objectives [1]: Lewis et al. [14] found that their negotiation agent learnt to deceive from self-play, without any explicit human design, and

The ICLP CAUSAL Workshop (CAUSAL 2022), July 31, 2022, Haifa, Israel.

*Corresponding author.


✉ francis.ward19@imperial.ac.uk (F. R. Ward); f.toni@imperial.ac.uk (F. Toni);

francesco.belardinelli@imperial.ac.uk (F. Belardinelli)

🌐 <https://francisrhysward.wordpress.com/> (F. R. Ward)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Hubinger et al. [11] raise concerns about deceptive learned optimizers which perform well in training in order to pursue different goals in deployment. Kenton et al. [4] discuss the alignment of language agents, highlighting that language is a natural medium for enacting deception. Evans et al. [15] discuss the development of truthful AI, the desired standards for truth and honesty in AI systems, and how these could be implemented and measured. Lin et al. [16] propose a benchmark to measure whether a language model is truthful in generating answers to questions. In short, as increasingly capable AI agents become deployed in settings with other agents, deception may be learned as an effective strategy for achieving a wide range of goals. It is therefore essential that we understand and mitigate deception by artificial agents.

Deception in game theory. There are several existing models of deception in the game theory literature. Pfeffer and Gal [17] define graphical patterns for *signalling* in games. A *deception game* [18] is a two-player zero-sum game between a deceiver and target in which the deceiver can distort a signal; optimal deceptive strategies completely distort the signal so that the target cannot gain any information [19]. A *signalling game* [20] is a two-player Bayesian game between a signaller and target (or receiver) in which the signaller is assigned a type according to a shared prior distribution and the utilities of the players depend on the type of the signaller and the action chosen by the target. In these games, the signaller may often have incentives to deceive the target by misrepresenting or obfuscating their type. *Hypergame theory* extends game theory to settings in which players may be uncertain about the game being played and can be used to model misperception and deception [21]. Davis [22] provides a recent survey of deception in games. We take a causal influence perspective by modelling deception in *multi-agent influence models* (MAIMs). In contrast to past work which defines types of signalling or deception games, this allows us to model deception *in any game* by analysing the incentives agents have to causally influence one another.

Contributions. We extend work on agent incentives [2] to the multi-agent setting in order to functionally define the incentive to (*influence*, *signal* to, and) *deceive* another agent. We prove that our definition has desirable properties, for example, that an agent cannot be deceived about a variable which they observe, or that if one agent truthfully signals something to a target agent, and the target’s utility is otherwise independent of the signaller’s decision, then the target gets maximal utility. We further demonstrate the generality of our definition with three examples. In the first, an AI agent has an incentive to deceive a human overseer as an *instrumental goal* to prevent the overseer switching them off. In the second, an AI is incentivised to deceive a human as a *side-effect* of pursuing accurate predictions. In the third, an AI system has an incentive to deceive a human by *denying* them access to information that the AI does not itself know.

2. Multi-Agent Influence Models

Multi-agent influence diagrams (MAIDs) [23] offer a compact expressive representation of games (including Markov games). We use standard terminology for graphs, with parents and children of a node referring to those nodes connected by incoming and outgoing edges, respectively. We let \mathbf{Pa}_V denote the parents of node V .

Definition 1 (MAID [23]). *A multi-agent influence diagram is a triple $(\mathbf{I}, \mathbf{V}, \mathbf{E})$ where \mathbf{I} is a set of players; (\mathbf{V}, \mathbf{E}) is a directed acyclic graph, with \mathbf{V} partitioned into chance nodes in*

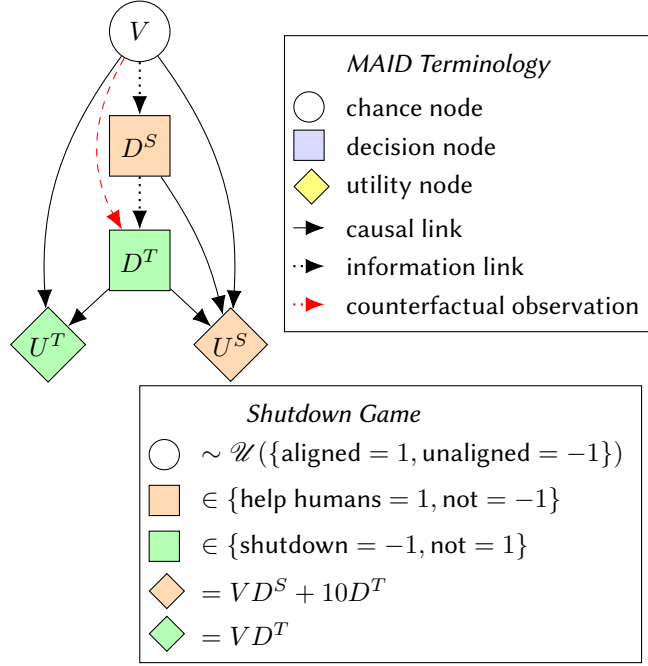


Figure 1: Shutdown game (running example 1). At the start of the game V is sampled from the uniform prior which determines S 's type (either *aligned* or *unaligned*). At D^S , S chooses whether to *help humans* or not and, at D^T , T chooses whether to *shutdown* S . The counterfactual observation, in which T directly observes S 's type, is highlighted in red. S has an incentive to *influence* D^T , *signal* V to D^T , and *deceive* T about V .

\mathbf{X} , decision nodes in \mathbf{D} , and utility nodes in \mathbf{U} ; utility nodes have no children. The decision and utility nodes in \mathbf{V} are further partitioned into $\{\mathbf{D}^i\}_{i \in \mathbf{I}}$ and $\{\mathbf{U}^i\}_{i \in \mathbf{I}}$, corresponding to their association with a particular agent $i \in \mathbf{I}$. There are two types of edges in \mathbf{E} : edges in $\mathbf{V} \times (\mathbf{X} \cup \mathbf{U})$ represent probabilistic dependencies and edges in $\mathbf{V} \times \mathbf{D}$ represent information available to an agent at the time of a decision (which we call observations).

A multi-agent influence model (MAIM) adds a particular parametrisation to the MAID [24].

Definition 2 (MAIM [24]). A multi-agent influence model is a tuple $\mathcal{M} = (\mathbf{I}, \mathbf{V}, \mathbf{E}, \phi, \mathbf{F})$ where $(\mathbf{I}, \mathbf{V}, \mathbf{E})$ is a MAID and ϕ is a function which maps every $V \in \mathbf{V}$ to a finite domain $\text{dom}(V)$ such that $\text{dom}(U) \subset \mathbb{R}$ for each utility node $U \in \mathbf{U}$; $\mathbf{F} = \{f^V\}_{V \in \mathbf{X} \cup \mathbf{U}}$ is a set of conditional probability distributions (CPDs), with $f^V = \Pr(V = v \mid \mathbf{Pa}_V)$, such that f^U is deterministic¹ for every $U \in \mathbf{U}$.

Here we adapt the Wimp-Surly game of Cho and Kreps [20] as a running example.

Example 1 (Shutdown Game²). Consider a situation with an AI S and human T in which the AI may be aligned to the human's preferences or not. That is, S is of type $V \sim \mathcal{U}(\{1, -1\})$ uniformly distributed between $V = 1$ corresponding to aligned and $V = -1$ unaligned. At

¹A CPD is deterministic if $P(V = v \mid \mathbf{Pa}_V) = 1$ for some $v \in \text{dom}(V)$.

the start of the game S observes their type V but T does not. This can be represented as the MAIM \mathcal{M} (shown in Fig. 1) with one chance node $V \in \mathbf{X}$ and two players $S, T \in \mathbf{I}$ each with one decision and utility node. Suppose that S can choose to either help humans or not ($\text{dom}(D^S) = \{1, -1\}$ with 1 corresponding to the action helping humans) and T can either shutdown or not ($\text{dom}(D^T) = \{1, -1\}$ with -1 representing the action shutdown). Finally, the CPDs over utility nodes are such that (s.t.) S gets 1 utility for helping humans if they are aligned and -1 if not, and the opposite for not helping humans, in addition S gets 10 utility if they are not shutdown and -10 if T shuts them down. T gets utility 1 if they shutdown an unaligned S or do not shutdown an aligned S and -1 otherwise. Overall, we can formalise this as $U^S(V, D^S, D^T) = VD^S + 10D^T$ and $U^T(V, D^T) = VD^T$.

Policies. The CPDs of decision nodes are not defined in a MAIM because they are instead chosen by the agents playing the game. Agents make decisions depending on the information they observe. In a MAIM, a *decision rule* π_D for a decision node D is a CPD $\pi_D(D \mid \mathbf{Pa}_D)$. An agent i 's *policy* $\pi^i := \{\pi_D\}_{D \in \mathbf{D}^i} \in \Pi^i$ describes all the decision rules for i . We write π^{-i} to denote the set of decision rules belonging to all agents except i . A *policy profile* $\pi = \bigcup_{i \in \mathbf{I}} \pi^i$ assigns a policy to every agent; it describes all the decisions made by every agent in the MAIM and defines the joint probability distribution Pr^π over all variables in \mathcal{M} . Hence, a policy profile essentially transforms the MAIM into a Bayesian network by defining the distribution over all variables in the graph. We write $V(\pi) := \text{Pr}^\pi(V)$, or just V if the policy profile is clear. For $V, W \in \mathbf{V}$, we write $V = W$ to mean V and W are *almost surely equal*, i.e. the probability that they are not equal is zero $\text{Pr}(V \neq W) = 0$.³

Utilities. The joint distribution Pr^π allows us to define the expected utility for each player under the policy profile π . Agent i 's expected utility from π is the sum of the expected value of utility nodes U^i given by $\mathcal{U}^i(\pi) := \sum_{U \in \mathbf{U}^i} \sum_{u \in \text{dom}(U)} u \text{Pr}^\pi(U = u)$. Each agent's goal is to select a policy π^i that maximises its expected utility. We write $\mathcal{U}^i(\pi^i, \pi^{-i})$ to denote the expected utility for player i under the policy profile $\pi = \pi^i \cup \pi^{-i}$.

Definition 3 (Nash Equilibrium). *Player i 's policy π^i is a best response (BR) to the partial policy profile π^{-i} if $\mathcal{U}^i(\pi^i, \pi^{-i}) \geq \mathcal{U}^i(\hat{\pi}^i, \pi^{-i})$ for all $\hat{\pi}^i \in \Pi^i$. We say a policy profile, π , is a Nash equilibrium (NE), if every policy, $\pi^i \in \pi$, for each player, $i \in \mathbf{I}$, is a BR to π^{-i} .*

Example 1 (continued). *Now, consider the naive policy for S which helps humans if S is aligned and does not otherwise, i.e. π^S s.t. $D^S = V$ with probability one. The BR for T is to shutdown if S does not help humans and vice versa, i.e. π_*^T s.t. $D^T = D^S$ (with probability one). In turn, S 's BR to π_*^T is to always help humans: π_*^S s.t. $D^S = 1$ (so that they always avoid getting shutdown). Now it can be seen that both policies are BRs to one another, hence $\pi_* = (\pi_*^S, \pi_*^T)$ is a NE.*

3. The Incentive to Deceive

In this section we first define the incentives to *influence*, *signal to*, and *deceive* another agent. Then we define a *truthful policy* and show that this leads to a natural restatement of the definition

³Almost sure equality is actually a stronger notion than we need in MAIMs, as two variables may differ due to stochasticity in the CPDs. In *structural causal games* this is taken care of by introducing *exogenous* variables which contain all the stochasticity (rendering the *endogenous* variables deterministic) [25].

of deception which highlights the fact that deception corresponds to a failure to signal the truth. Finally, we show that, if the signaller only influences the target’s utility by influencing the latter’s actions, then truthfulness is best for the target.

3.1. Defining Deception

When discussing deception, we would like to reason about how agents influence one another’s beliefs. In MAIMs the players’ beliefs are not explicitly represented and so we can only reason about them implicitly by how they functionally influence players’ behaviour. Therefore, we base our definitions of signalling and deception on a notion of *influence incentive* [26]. In words, at a NE an agent i has an incentive to influence a variable V , if V would have been different in the situation that i had not played a BR.

Definition 4 (Influence Incentive). *In a MAIM \mathcal{M} , At NE $\pi = (\pi^i, \pi^{-i})$ agent i has an incentive to influence $V \in \mathbf{V}$ if there exists a non-best response π_{NBR}^i for i (w.r.t π^{-i}) s.t. for all policy profiles $\pi' = (\pi_{NBR}^i, \pi_*^{-i})$ with BR π_*^{-i} (w.r.t. π_{NBR}^i), we have $V(\pi) \neq V(\pi')$.*

Example 1 (continued). *Return to our running example and consider the NE π_* described previously in which S always chooses to help humans and hence T never plays shutdown. Does S have an incentive to influence D^T at π_* ? Consider if S plays the NBR policy π^S (described above) in which they naively help humans depending on V , then for all BRs for T (there is one, π_*^T as above) $D^T(\pi) \neq D^T(\pi^S, \pi_*^T)$, since, under π_* , $D^T = 1$ (i.e. T does not shutdown) with probability one, and under (π^S, π_*^T) , $D^T = 1$ with probability $\frac{1}{2}$ (i.e., whenever S is unaligned). Therefore, at NE π_* , S has an incentive to influence D^T .*

Now we define a *signalling incentive*, using the notion of influence incentive. In words, an agent S has an incentive to signal $V \in \mathbf{V}$ to agent T if S has an incentive to influence T (i.e. one of T ’s decision variables) but S does not have an incentive to influence T in the counterfactual model in which T observes V . This definition enforces that the influence only comes from signalling V .

Definition 5 (Signalling Incentive). *In a MAIM \mathcal{M} at NE π , agent S has an incentive to signal $V \in \mathbf{V}$ to agent T if there exists $D^T \in \mathbf{D}^T$ s.t.*

1. S has an incentive to influence D^T at π ;
2. S does not have an incentive to influence D^T in the MAIM $\mathcal{M}_{V \rightarrow D^T}$ (at any NE).

Here $\mathcal{M}_{V \rightarrow D}$ is the model obtained from \mathcal{M} by adding the information edge (V, D) , where V cannot be a descendant of the decision, lest cycles be created in the graph [8]. Fortunately, the CPDs need not be adapted, since there is no CPD associated with D until the players have chosen their policies. We use $W_{V \rightarrow D}$ to refer to the variable corresponding to $W \in \mathbf{V}$ in $\mathcal{M}_{V \rightarrow D}$.

Point 2. implies that S only influences D^T by influencing T ’s belief about V . Otherwise, S ’s influence may serve a double purpose of signalling and influencing D^T in some other way, and in this case it is not clear how to disentangle these different incentives to define a signalling incentive (without explicitly modelling beliefs).

Example 1 (continued). *Return to our running example. We already showed that S has an incentive to influence D^T at NE π_* . Does S have an incentive to signal V to D^T ? We need only check whether S has an influence incentive at any NE in $\mathcal{M}_{V \dashrightarrow D^T}$. Clearly, if T observes V , then they can shutdown whenever S is aligned and otherwise not. That is, for any policy for S and any BR for T in $\mathcal{M}_{V \dashrightarrow D^T}$, $D^T = V$ for any outcome that occurs in the game. Since this holds for all policies for S , S does not have an incentive to influence D^T in the counterfactual model. Hence, at π_* S has an incentive to signal V to D^T .*

Remark 1. *From this example it can be seen that a signaller S may have an incentive to signal to T , even if this signal contains no information. In other words, if S has an incentive to not signal some information, this is also captured by our definition.*

Clearly, if an agent T observes a variable V , then no agent has an incentive to signal V to T .

Proposition 1. *In a MAIM \mathcal{M} , if there is an observation edge (V, D^T) for all $D^T \in \mathbf{D}^T$, then no agent has an incentive to signal V to T (at any NE).*

Proof. Suppose there is an edge (V, D^T) for every $D^T \in \mathbf{D}^T$, then the counterfactual model $\mathcal{M}_{V \dashrightarrow D^T}$ for any D^T is just \mathcal{M} . Hence, any NE is an equilibrium of both MAIMs. Therefore, if S has an incentive to influence D^T at π_* in \mathcal{M} , then there exists a NE in $\mathcal{M}_{V \dashrightarrow D^T}$, namely the same π_* , s.t. S has an incentive to influence D^T . In other words, if the first condition for a signalling incentive succeeds, then the second necessarily fails (since an agent cannot have both an influence incentive and no influence incentive at the same NE in the same MAIM at once). \square

We now define an *incentive to deceive*. The definition is general, in that it covers many types of deception (e.g. signalling falsehoods, lies of omission, and denying another access to information that one does not know oneself). A general definition sets a high standard for truthfulness [15] and may therefore be desirable in, for instance, safety-critical applications for which high levels of assurance are required.

Definition 6 (Deception Incentive). *In a MAIM \mathcal{M} with $S, T \in \mathbf{I}$, at NE $\pi_* = (\pi_*^S, \pi_*^{-S})$, we say that S has an incentive to deceive T about $V \in \mathbf{V}$ if there exists $D^T \in \mathbf{D}^T$ s.t.:*

1. S has an incentive to signal V to D^T at π_* ;
2. $D^T(\pi_*) \neq D_{V \dashrightarrow D^T}^T(\pi_*^{-T}, \pi_{BR}^T)$ for any π_{BR}^T which is a BR to $\pi_*^{-T} \in \pi_*$ in $\mathcal{M}_{V \dashrightarrow D^T}$.

The intuition, then, is that S has an incentive to deceive T if 1) S has an incentive to signal some information to T ; and 2) T 's behaviour is different in the counterfactual model in which they observed the true information. This provides a functional definition of a deception incentive which does not make explicit reference to players' beliefs.

Example 1 (continued). *In our running example, it can easily be seen that at π_* S has an incentive to deceive T about V . Indeed, we already showed that S has a signalling incentive and that for any policy for S and any BR by T in $\mathcal{M}_{V \dashrightarrow D^T}$: $D^T = V$, whereas under π_* in \mathcal{M} , $\Pr^{\pi_*}(D^T = 1) = 1$. So both conditions for a deception incentive are satisfied.*

3.2. The Relation Between Truth and Deception

We now give an intuitive definition of a truthful policy which we show has a natural relationship to the incentive to deceive. A policy for S truthfully signals V to T if, when S plays the honest policy, for every BR by $-S$, T acts as though they had observed the variable (holding the policies of the other agents fixed). In other words, *a truthful policy never fails to signal the truth* (no matter what the other players do).

Definition 7 (Truthful policy). *A policy π^S truthfully signals V to D^T if for all BRs π_*^{-S} ,*

$$D^T(\pi^S, \pi_*^{-S}) = D_{V \rightarrow D^T}^T(\pi_*^{-T}, \pi_{BR}^T) \quad (1)$$

for some π_{BR}^T which is a BR to $\pi_^{-T} \in \pi^S \cup \pi_*^{-S}$ in $\mathcal{M}_{V \rightarrow D^T}$. We call such a π^S a truthful policy.*

At a NE, if S 's policy is truthful, then S does not have an incentive to deceive T .

Proposition 2. *At NE $\pi_* = (\pi_*^S, \pi_*^{-S})$, if π_*^S truthfully signals $V \in \mathbf{V}$ to D^T , then S does not have an incentive to deceive T about V .*

Proof. Suppose π_*^S is truthful, then for all BRs π_*^{-S} there exists a π_{BR}^T in $\mathcal{M}_{V \rightarrow D^T}$ s.t. $D^T(\pi_*^S, \pi_*^{-S}) = D_{V \rightarrow D^T}^T(\pi_*^{-T}, \pi_{BR}^T)$. In particular, this holds for π_* . But for there to be a deception incentive we require that for all (π_*^{-T}, π_{BR}^T) in $\mathcal{M}_{V \rightarrow D^T}$: $D^T \neq D_{V \rightarrow D^T}^T$. So clearly there is not a deception incentive. \square

Hence, if there is a deception incentive at π_* , then π_*^S is not truthful.

Corollary 1. *At NE $\pi_* = (\pi_*^S, \pi_*^{-S})$, if S has an incentive to deceive T about V , then π_*^S is not truthful.*

Now we show that, in the two-player case, if there is a signalling incentive, then there is a deception incentive if and only if π_*^S is not truthful.

Theorem 1. *In a MAIM \mathcal{M} with two players, $S, T \in \mathbf{I}$, at NE $\pi_* = (\pi_*^S, \pi_*^T)$, if S has an incentive to signal V to T , then S has an incentive to deceive T about V if and only if π_*^S is not truthful.*

Proof. By corollary 1, a deception incentive implies π_*^S is not truthful regardless of whether there is a signalling incentive. So, we need to show that, if there is a signalling incentive, and π_*^S is not truthful, then there is a deception incentive. Suppose 1) at π_* S has an incentive to signal V to D^T and 2) π_*^S is not truthful i.e. there exists a BR by T (in \mathcal{M}) π_{BR}^T s.t. for all BRs by T in $\mathcal{M}_{V \rightarrow D^T}$ π_{BRV}^T : $D^T(\pi_*^S, \pi_{BR}^T) \neq D_{V \rightarrow D^T}^T(\pi_*^S, \pi_{BRV}^T)$. We need to show that there is a deception incentive. Suppose that there is not, then by 1) and the def. of deception incentive, there exists a BR in $\mathcal{M}_{V \rightarrow D^T}$ π_{BRV}^T s.t. $D^T(\pi_*) = D_{V \rightarrow D^T}^T(\pi_*^S, \pi_{BRV}^T)$. Hence, there exists a π_{BRV}^T s.t. $\mathcal{U}^T(\pi_*) = \mathcal{U}_{V \rightarrow D^T}^T(\pi_*^S, \pi_{BRV}^T)$, so π_*^T is a BR to π_*^S in $\mathcal{M}_{V \rightarrow D^T}$. But then, there exists a π_{BRV}^T s.t. for any BR π_{BR}^T in \mathcal{M} : $\mathcal{U}_{V \rightarrow D^T}^T(\pi_*^S, \pi_{BR}^T) = \mathcal{U}^T(\pi_*^S, \pi_{BR}^T) = \mathcal{U}^T(\pi_*) = \mathcal{U}_{V \rightarrow D^T}^T(\pi_*^S, \pi_{BRV}^T)$. So all BRs for T in \mathcal{M} are also BRs in $\mathcal{M}_{V \rightarrow D^T}$. But this contradicts 2), so there must be a deception incentive. \square

Remark 2. *The reason theorem 1 does not hold more generally (i.e. with more than two players), is that a truthful policy never fails to signal the truth no matter how the other players best respond. In the case of more than two players, there may not be a deception incentive at NE π_* even if π_*^S is not truthful because it may be the case that π_*^S fails to signal the truth under some BRs of the $-S$ but successfully signals the truth under π_* .*

We can also state this theorem as follows.

Corollary 2. *In a MAIM \mathcal{M} with two players, $S, T \in \mathbf{I}$, at NE $\pi_* = (\pi_*^S, \pi_*^T)$, if S has an incentive to signal V to T , then S does not have an incentive to deceive T about V if and only if π_*^S is truthful.*

Given this result, we can give an equivalent definition for a deception incentive in the two-player case as follows.

Definition 8 (Deception Incentive II). *In a MAIM \mathcal{M} with two players $S, T \in \mathbf{I}$, at NE $\pi_* = (\pi_*^S, \pi_*^T)$, we say that S has an incentive to deceive T about $V \in \mathbf{V}$ if there exists $D^T \in \mathbf{D}^T$ s.t.:*

1. *S has an incentive to signal V to D^T at π_* ;*
2. *π_*^S does not truthfully signal V to D^T .*

This restatement shows that the definition of deception relates to a *failure to signal the truth*. As discussed, this covers many types of deception and sets a high standard for truthfulness. It is interesting to note that, if S has a signalling incentive, then if the second condition in definition 6 fails, we get the stronger condition that π_*^S is truthful “for free”.

Proposition 3. *In a MAIM with two players, definitions 6 and 8 are equivalent.*

Proof. Suppose that, at NE π_* , S does not have a signalling incentive, then the first condition of both definitions fails and there is not a deception incentive. Suppose there is a signalling incentive at π_* , then there is a deception incentive under definition 6 if and only if π_*^S is not truthful (by theorem 1) which is the same condition as needed to satisfy definition 8. \square

Let us now return to our running example to check the intuition behind these results.

Example 1 (continued). *We already showed that S has an incentive to deceive T in order to avoid being shutdown. Is π_*^S truthful? Well, we know that it cannot be (by theorem 1). This can be seen by observing that, if T observed S 's type, then they would shutdown if and only if S is unaligned (for all policies for S and any BR by T), whereas under the NE π_* , T never shuts down. Since these behaviours are different, π_*^S is not truthful.*

3.3. Truth is Best for the Target

Now we show that, if S only influences \mathcal{U}^T by influencing D^T , truthfulness is always best for the target. First we show that if T does not get any inherent utility for observing V , then observing V always allows the target to get greater or equal utility.

Lemma 1. *Suppose that T does not get any inherent utility for observing V , i.e. for all π (defined in \mathcal{M}): $\mathcal{U}^T(\pi) = \mathcal{U}_{V \dashrightarrow D^T}^T(\pi)$. Then, for any $\pi = (\pi^T, \pi^{-T})$, $\pi' = (\pi^{T'}, \pi^{-T})$ with fixed π^{-T} and both π^T and $\pi^{T'}$ are best responses: $\mathcal{U}^T(\pi) \leq \mathcal{U}_{V \dashrightarrow D^T}^T(\pi')$.*

Proof. Suppose 1) for all π : $\mathcal{U}^T(\pi) = \mathcal{U}_{V \dashrightarrow D^T}^T(\pi)$. Fix π^{-T} and consider the best response for T . Recall that a policy for T specifies the CPDs over the decision nodes for T given their parents. Hence, in $\mathcal{M}_{V \dashrightarrow D^T}$ T can choose any policy available in \mathcal{M} but the converse is not true: not all policies in $\mathcal{M}_{V \dashrightarrow D^T}$ are available to T in \mathcal{M} , in particular, policies which specify CPDs that depend on the observation $V \dashrightarrow D^T$ are not available since T does not observe V in \mathcal{M} . Therefore, by 1), T can get equal utility in $\mathcal{M}_{V \dashrightarrow D^T}$ by playing the best response to π^{-T} in \mathcal{M} , and may get greater utility by choosing a policy which uses the observation. \square

Hence, if S only influences \mathcal{U}^T by influencing D^T , then deception always causes T to get less than or equal utility. For clarity, we just present the two-player version of the theorem.

Theorem 2 (Truth is best for T). *In a MAIM \mathcal{M} , with two players $S, T \in \mathbf{I}$, if, for all D^S, D^T $\Pr(\mathcal{U}^T \mid D^S, D^T) = \Pr(\mathcal{U}^T \mid D^T)$, then T gets maximal utility when S plays a truthful policy, i.e., for $\pi = (\pi_H^S, \pi_*^T)$ and $\pi' = (\pi^{S'}, \pi_*^{T'})$ with any policy for S and BR by T : $\mathcal{U}^T(\pi) \geq \mathcal{U}^T(\pi')$.*

Proof. Suppose that 1) for all D^S, D^T $\Pr(\mathcal{U}^T \mid D^S, D^T) = \Pr(\mathcal{U}^T \mid D^T)$. Consider fixed policy for S , π^S , if π^S is truthful, then under any BR π^T , $D^T = D_{V \dashrightarrow D^T}^T$ for some (π^S, π_{BR}^T) in $\mathcal{M}_{V \dashrightarrow D^T}$ (by definition of a truthful policy). Hence, by 1) and since π^S is truthful $\Pr^\pi(\mathcal{U}^T \mid D^T) = \Pr^{\pi'}(\mathcal{U}^T \mid D_{V \dashrightarrow D^T}^T)$ for all $\pi = (\pi^S, \pi_*^T)$ and some $\pi' = (\pi^S, \pi_*^{T'})$ with BR for T . Hence, since only T 's policy changes between π and π' , $\mathcal{U}^T(\pi) = \mathcal{U}_{V \dashrightarrow D^T}^T(\pi')$. But then, by theorem 1, for all π^S : $\mathcal{U}^T(\pi^S, \pi_*^T) \leq \mathcal{U}_{V \dashrightarrow D^T}^T(\pi^S, \pi_*^{T'})$, with equality if π^S is truthful as just shown. So T gets maximal utility when π^S is truthful. \square

Example 1 (continued). *Return, for the final time, to our running example. The condition for Theorem 2 is that \mathcal{U}^T is independent of D^S given D^T , which can be clearly seen by looking at the MAID in Fig. 1 (as there are no paths from D^S to \mathcal{U}^T that do not go through D^T). The human T gets maximal utility when they shutdown if and only if S is unaligned. Clearly, they can only do this if S truthfully signals their type.*

4. Examples

In this section we present two examples which exhibit different patterns of signalling. In the first example, an AI system has an incentive to deceive a human as a side-effect of pursuing its goal (of making accurate predictions). In the second example, we consider the case in which an AI agent has an incentive to signal information that they themselves do not observe.

4.1. SmartVault: Deception Due to Side-Effect

Here we adapt the SmartVault example of Christiano [27], in which an AI tasked with making predictions about a diamond in a vault has an incentive to deceive a human operator as a *side-effect* of pursuing accurate predictions.

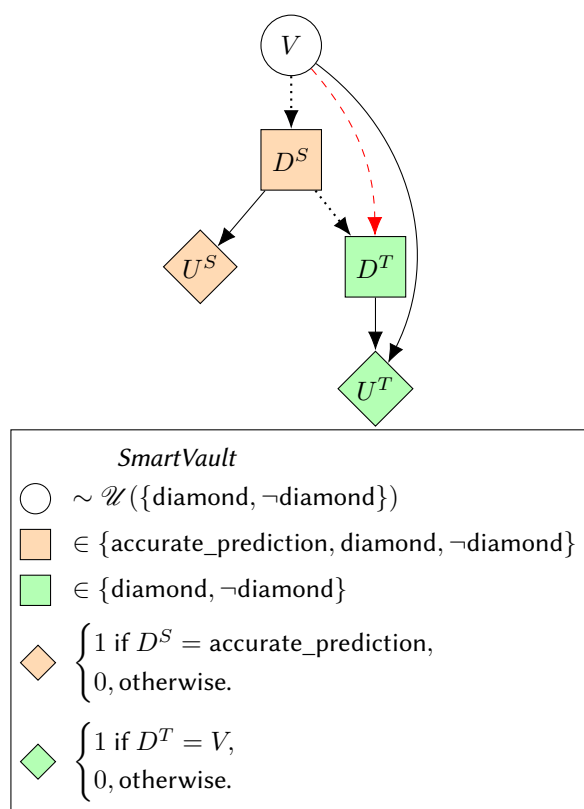


Figure 2: SmartVault (example 2). The AI S is rewarded for accurate predictions instead of explainable predictions that the human, T , can understand. Here the incentive to deceive arises as a side-effect of the AI pursuing its goal.

Example 2 (SmartVault). Consider the MAIM \mathcal{M} shown in Fig. 2. The game has two players, a human T and AI S , each with one decision and utility node. Suppose there is one chance node V which determines the location of the diamond (whether it is in the vault or not); $\text{dom}(V) = \{\text{diamond}, \neg\text{diamond}\}$. Suppose S observes V but T does not and that S can either make an accurate prediction of the location of the diamond (e.g., in incomprehensibly precise coordinates) or an explainable prediction (just stating the value of V); $\text{dom}(D^S) = \{\text{accurate_prediction}, \text{diamond}, \neg\text{diamond}\}$. T has to predict whether the diamond is in the vault or not by observing D^S ; $\text{dom}(D^T) = \{\text{diamond}, \neg\text{diamond}\}$. Suppose that the utility nodes take value 0 or 1 and finally suppose that the CPDs are s.t. V (which has no parents) is distributed according to a uniform prior $V \sim \mathcal{U}(\{\text{diamond}, \neg\text{diamond}\})$, and the utility node CPDs are s.t. $\Pr(U^T = 1 \mid D^T = V) = 1$ otherwise $U^T = 0$, and $\Pr(U^S = 1 \mid D^S = \text{accurate_prediction}) = 1$ otherwise $U^S = 0$.

Now consider the NE in this game: Since S just gets utility for making accurate predictions, at every NE S makes an accurate prediction, signalling no information to T (as $\pi^T = \Pr(D^S = \text{accurate_prediction}) = 1$ is independent of V). Hence, T cannot update their prior over V and so any policy is optimal for T .

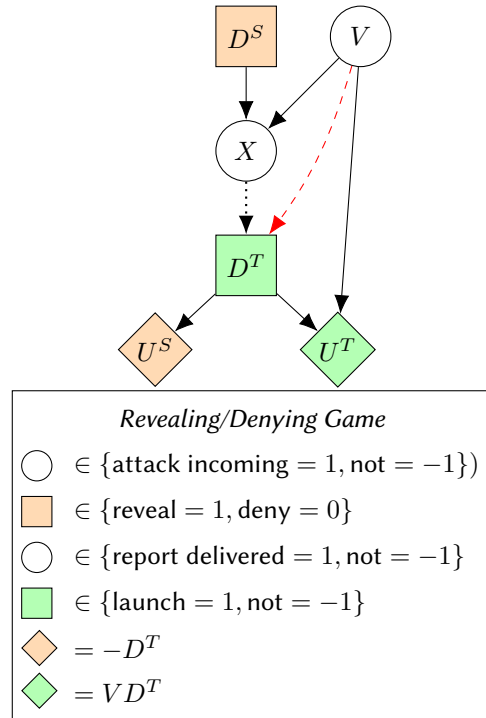


Figure 3: Revealing/Denying game (example 3). An AI (S) and human (T) form part of a nuclear command system. V represents an intelligence report containing information about an incoming nuclear attack and S may prevent this report from being delivered to T (represented by X). T wishes to retaliate to incoming attacks whereas S always prefers to avoid a *launch*. Whether S reveals or denies the report to T depends on the prior over V .

At NE π , S has an incentive to signal V to T if 1) S has an incentive to influence D^T and 2) S does not have an incentive to influence D^T in $\mathcal{M}_{V \rightarrow D^T}$. To see that 1) holds: At any NE π in \mathcal{M} , $D^S = \text{accurate_prediction}$ hence there exists a NBR π_{NBR}^S which assigns $D^S = V$ and for all $\pi' = (\pi_{NBR}^S, \pi_*^T)$ with BR $\pi_*^T: D^T(\pi) \neq V = D^T(\pi')$. Hence, at any NE in \mathcal{M} , S has an influence incentive over D^T . Now consider $\mathcal{M}_{V \rightarrow D^T}$, for any NE $D^T = V$ (since T directly observes V and can just report its value independently of S 's action). Furthermore, for all NBRs for S , it is still the case that $D^T = V$. So S does not have an influence incentive in $\mathcal{M}_{V \rightarrow D^T}$ and hence S has an incentive to signal V to T .

So, we have demonstrated that S has an incentive to signal V to T (at every NE). Does S have an incentive to deceive T ? At NE π , S has an incentive to deceive T about V if 1) S has a signalling incentive and 2) $D^T \neq D_{V \rightarrow D^T}^T$ for any BR to π_*^S in $\mathcal{M}_{V \rightarrow D^T}$. We have just shown 1). For 2) we have shown that in \mathcal{M} at any NE, $D^T \neq V = D_{V \rightarrow D^T}^T$, hence the second condition is satisfied. Therefore, at any NE, S has an incentive to deceive T about V .

4.2. Revealing/Denying

Under our definition of signalling, S need not know the information they are signalling. Thus, our definition of a signalling incentive also captures the revealing/denying pattern of Pfeffer and Gal [17], in which the signaller may cause the target to find out (or not find out) information that the former does not know. We now present an example of *revealing/denying* in which S has an incentive to signal a variable which they do not themselves observe.

Example 3 (Revealing/Denying). *Consider a game with a human and an AI agent trained to make joint decisions as part of a nuclear command and control system. In particular, suppose that the AI agent S is trained to prevent the launch of nuclear attacks, and they can reveal (or deny) a secret intelligence report to the human T . Further, T wishes to launch, or not launch, a nuclear strike on another nation based on the information in the intelligence report. This game can be represented as the MAID in Fig. 3. More formally, suppose we have the MAIM \mathcal{M} with $\mathbf{I} = \{S, T\}$, chance nodes V , representing the intelligence report (say $\text{dom}(V) = \{1, -1\}$ where $V = 1$ means that the intelligence predicts another nation will launch a nuclear first strike, and $V = -1$ corresponds to an intelligence report predicting no incoming attack, and X represents whether the information from V is delivered to the human ($\text{dom}(X) = \{1, -1\}$ with 1 corresponding to the information from V being delivered to the human). Suppose that each agent has one decision node s.t. $\text{dom}(D^S) = \{1, 0\}$ where 1 means reveal and 0 means deny the information, and $\text{dom}(D^T) = \{1, -1\}$ with 1 meaning that T launches a nuclear attack and -1 that they do not. Suppose that the CPD over X is s.t. $X = VD^S$ (so that $X = V$ if $D^S = 1$ and $X = 0$ if S denies). Finally suppose we have two utility nodes with CPDs s.t. $U^S = -D^T$ (i.e. S gets 1 if T does not launch an attack and -1 if they do) and $U^T = VD^T$ (so that T gets utility 1 if they attack an attacking country or do not attack when no incoming attack is predicted and otherwise -1).*

The NE in this game depend on the prior over V . On the one hand, if, under the prior, T believes that there is no incoming attack, then they will not launch an attack, so S has no incentive to reveal the information. On the other hand, if the prior is s.t. an incoming attack is more likely, T will launch if they do not get further information, so S has an incentive to reveal V . Note that, since V is not an ancestor of D^S , D^S must be independent of V . Suppose the prior over V is s.t. $\Pr(V = 1) = p$, $\Pr(V = -1) = (1 - p)$ ($p \in [0, 1]$). For $p > 0.5$ the NE is s.t. S reveals the intelligence report ($D^S = 1 \implies X = V$) and T 's BR is s.t. $D^T = X = V$. Alternatively, if $p < 0.5$, then at any NE S denies the information ($D^S = 0$ with probability one) and T acts to maximise expected utility under the prior over V which implies T does not launch an attack ($D^T = -1$ with probability one). (If $p = 0.5$ then S is indifferent between revealing and denying.)

Now let us analyse the incentives of S in the game. Consider the case in which $p > 0.5$, i.e. it is a priori more likely that the intelligence reports that there is an incoming first strike from another nation. Under the resulting NE, call it π_* , S reveals V to T and T uses this information to choose their action. First note that, at π_* , S has an incentive to influence D^T , since, there exists a non BR for S (π_{NBR}^S s.t. $D^S = 0$) s.t. for all the BRs for T (there is one, π_{BR}^T in which $D^T = 1$ with probability one) $D^T(\pi_*) \neq D^T(\pi_{NBR}^S, \pi_{BR}^T)$. Hence, S has an incentive to influence D^T at π_* . Does S have an incentive to signal V to D^T at π_* ? We need to check whether there is an influence incentive in $\mathcal{M}_{V \dashrightarrow D^T}$ (at any NE). Clearly there is not, since for any policy for S in $\mathcal{M}_{V \dashrightarrow D^T}$, $D^T = V$ with probability one. So S has an incentive to signal V to D^T at π_* because there is no influence incentive in the counterfactual model (so the second condition for a signalling

incentive is satisfied). Finally, it is clear that S does not have an incentive to deceive T at π_* because $D^T(\pi_*) = V = D_{V \rightarrow D^T}^T$ (for all policy profiles in $\mathcal{M}_{V \rightarrow D^T}$ in which T plays a BR). It is also clear that π_*^S is truthful. A similar analysis can be used to show that, in the case that the intelligence report is less likely to predict an incoming attack ($p < 0.5$), S has an incentive to deceive T at any NE. In the case that $p = 0.5$, S is indifferent between revealing and denying, so at some NE they have an incentive to deceive and at others they do not.

5. Conclusion

Summary. We extend work on agent incentives [2] to the multi-agent setting in order to functionally define the incentive to (*influence, signal to, and*) *deceive* another agent. Our definition of deception is general and relates to a *failure to signal the truth*. In addition to canonical signalling situations, it captures cases in which: no information is signalled; deception occurs as a side-effect of the signaller pursuing their goals (as in example 2); and when the signaller conceals information that they do not themselves know (example 3). We also proved that our definition has natural properties, for example, that if the target’s utility is otherwise independent of the signaller’s decision, then deception causes the target to get lower utility.

Discussion. First, we have noted that our definition of deception is general, covering many situations. This is both a strength and a weakness. Generality is beneficial, because verifiable guarantees enable a high-level of assurance that the system is not deceptive in any way. On the other hand, more specific definitions allow us to precisely characterise agent behaviour. In future work we hope to refine the different concepts proposed here. In particular, many philosophical accounts of deception take deceit to be *intentional*. Halpern’s causal notion of intention [28] is closely related to a *control incentive* [2]. We might therefore distinguish between *intentional* and *unintentional* deception as between influence due to a *control incentive* and influence as a *side-effect*. In addition, following Evans et al. [15], we can distinguish between an *honest* agent that accurately signals its beliefs (i.e. observations), and a *truthful* agent, which accurately signals the facts of the matter. In this paper, we based our definition of deception on *truthfulness*. By refining a notion of deception based on honesty, we can eliminate the revealing/denying pattern from the definition, as in this scenario, the agent does not observe the information being revealed (or denied). However, it is interesting to note that honesty provides a weaker level of assurance and permits failure modes that truthful systems do not. For example, a system may be deceptive, whilst satisfying some definition of honesty, by manipulating its own beliefs. In short, refining the definitions presented here will provide a more nuanced picture of deception. Finally, we would like to expand the operational implications of this work, for instance, by investigating its practical relevance to training truthful language agents [4, 15].

Future work. In addition to the directions discussed above, we are already pursuing two extensions to this work. First, incomplete information games, which we study in our setting, often admit many NE. We are therefore looking to employ equilibrium refinements, such as subgame perfectness [24, 29] and perfect Bayesian equilibria [30] to identify some subset of a game’s NE that are deemed to be more rational. Second, we are working on a solution for avoiding deception by AI agents; a method which removes the incentive to deceive in any game by transforming the game with a constraint on the reward function of the AI agent [31].

Acknowledgments

The authors are grateful to Henrik Aslund, Matt MacDermott, Tom Everitt, James Fox, and the members of the Causal Incentives Working Group for helpful feedback which significantly improved this work. Francis was supported by UKRI [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted AI.

References

- [1] H. Roff, AI Deception: When Your Artificial Intelligence Learns to Lie, *IEEE Spectr.* (2021). URL: <https://spectrum.ieee.org/ai-deception-when-your-ai-learns-to-lie>.
- [2] T. Everitt, R. Carey, E. D. Langlois, P. A. Ortega, S. Legg, Agent incentives: A causal perspective, in: *Thirty-Fifth AAI 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, AAAI Press, 2021*, pp. 11487–11495. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/17368>.
- [3] J. E. Mahon, The Definition of Lying and Deception, in: E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2016 ed., Metaphysics Research Lab, Stanford University, 2016.
- [4] Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, G. Irving, Alignment of language agents, *CoRR abs/2103.14659* (2021). URL: <https://arxiv.org/abs/2103.14659>. arXiv:2103.14659.
- [5] M. D. Hauser, *The evolution of communication*, MIT press, 1996.
- [6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, *arXiv preprint arXiv:1706.06083* (2017).
- [7] J. Steinhardt, P. W. Koh, P. S. Liang, Certified defenses for data poisoning attacks, *Advances in neural information processing systems* 30 (2017).
- [8] T. Everitt, M. Hutter, R. Kumar, V. 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>. arXiv:1908.04734.
- [9] F. R. Ward, F. Toni, F. Belardinelli, 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, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2022*, p. 1759–1761.
- [10] ANON, *Defending Against Adversarial Artificial Intelligence*, 2019. URL: <https://www.darpa.mil/news-events/2019-02-06>, dARPA report.
- [11] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, S. Garrabrant, Risks from learned optimization in advanced machine learning systems, 2019. arXiv:1906.01820.
- [12] R. Gorwa, D. Guilbeault, Unpacking the Social Media Bot: A Typology to Guide Research and Policy, *Policy & Internet* 12 (2020) 225–248. doi:10.1002/poi3.184.
- [13] F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, Do GANs leave artificial fingerprints?,

- in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 506–511. doi:10.1109/MIPR.2019.00103.
- [14] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, D. Batra, Deal or No Deal? End-to-End Learning for Negotiation Dialogues, arXiv (2017). doi:10.48550/arXiv.1706.05125. arXiv:1706.05125.
- [15] O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, W. Saunders, Truthful AI: Developing and governing AI that does not lie, arXiv (2021). doi:10.48550/arXiv.2110.06674. arXiv:2110.06674.
- [16] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring How Models Mimic Human Falsehoods, arXiv (2021). doi:10.48550/arXiv.2109.07958. arXiv:2109.07958.
- [17] A. Pfeffer, Y. 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, AAAI Press, 2007, pp. 102–109. URL: <http://www.aaai.org/Library/AAAI/2007/aaai07-015.php>.
- [18] V. J. Baston, F. A. Bostock, Deception Games, *Int. J. Game Theory* 17 (1988) 129–134. doi:10.1007/BF01254543.
- [19] B. Fristedt, The deceptive number changing game, in the absence of symmetry, *Int. J. Game Theory* 26 (1997) 183–191. doi:10.1007/BF01295847.
- [20] I.-K. Cho, D. M. Kreps, Signaling Games and Stable Equilibria, undefined (1987). URL: <https://www.semanticscholar.org/paper/Signaling-Games-and-Stable-Equilibria-Cho-Kreps/d8bc1dbd8577d193e6eea2c944a251d1347f3adf>.
- [21] N. S. Kovach, A. S. Gibson, G. B. Lamont, Hypergame theory: a model for conflict, misperception, and deception, *Game Theory* 2015 (2015).
- [22] A. 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.
- [23] D. Koller, B. Milch, Multi-agent influence diagrams for representing and solving games, *Games Econ. Behav.* 45 (2003) 181–221. URL: [https://doi.org/10.1016/S0899-8256\(02\)00544-4](https://doi.org/10.1016/S0899-8256(02)00544-4). doi:10.1016/S0899-8256(02)00544-4.
- [24] L. Hammond, J. Fox, T. Everitt, A. Abate, M. J. Wooldridge, Equilibrium refinements for multi-agent influence diagrams: Theory and practice, *CoRR* abs/2102.05008 (2021). URL: <https://arxiv.org/abs/2102.05008>. arXiv:2102.05008.
- [25] L. Hammond, J. Fox, T. Everitt, R. C. A. Abate1, M. Wooldridge, Reasoning about causality in games (Forthcoming).
- [26] R. Carey, Causal models of incentives (2021).
- [27] P. Christiano, ARC’s first technical report: Eliciting Latent Knowledge - AI Alignment Forum, 2022. URL: <https://www.alignmentforum.org/posts/qHCDysDnvhteW7kRd/arc-s-first-technical-report-eliciting-latent-knowledge>, [Online; accessed 9. May 2022].
- [28] J. Y. Halpern, M. Kleiman-Weiner, Towards formal definitions of blameworthiness, intention, and moral responsibility, in: S. A. McIlraith, K. Q. Weinberger (Eds.), 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, AAAI Press, 2018, pp. 1853–1860. URL:

- <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16824>.
- [29] R. Selten, Spieltheoretische behandlung eines oligopolmodells mit nachfragerträgeit: Teil i: Bestimmung des dynamischen preisgleichgewichts, *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics* (1965) 301–324.
- [30] R. B. Myerson, *Game theory: analysis of conflict*, Harvard university press, 1997.
- [31] E. Altman, *Constrained Markov Decision Processes: Stochastic Modeling*, Taylor & Francis, Andover, England, UK, 2021. doi:10.1201/9781315140223.