Agent Incentives: A Causal Perspective

Tom Everitt,[*1*] **Ryan Carey,**[*2] **Eric D. Langlois,**[*1,3,4] **Pedro A. Ortega,**[1] **Shane Legg**[1]

[1]DeepMind, [2]University of Oxford, [3]University of Toronto, [4]Vector Institute

tomeveritt@google.com, ry.duff@gmail.com, edl@cs.toronto.edu, pedroortega@google.com

Abstract

We present a framework for analysing agent incentives using causal influence diagrams. We establish that a well-known criterion for value of information is complete. We propose a new graphical criterion for value of control, establishing its soundness and completeness. We also introduce two new concepts for incentive analysis: response incentives indicate which changes in the environment affect an optimal decision, while instrumental control incentives establish whether an agent can influence its utility via a variable X. For both new concepts, we provide sound and complete graphical criteria. We show by example how these results can help with evaluating the safety and fairness of an AI system.

Introduction

A recurring question in AI research is how to choose an objective to induce safe and fair behaviour (O'Neil 2016; Russell 2019). In a given setup, will an optimal policy depend on a sensitive attribute, or seek to influence an important variable? For example, consider the following two incentive design problems, to which we will return throughout the paper:

Example 1 (Grade prediction). To decide which applicants to admit, a university uses a model to predict the grades of new students. The university would like the system to predict accurately, without treating students differently based on their gender or race (see Figure 1a).

Example 2 (Content recommendation). An AI algorithm has the task of recommending a series of posts to a user. The designers want the algorithm to present content adapted to each user's interests to optimize clicks. However, they do not want the algorithm to use polarising content to manipulate the user into clicking more predictably (Figure 1b).

Contributions This paper provides a common language for incentive analysis, based on influence diagrams (Howard 1990) and causal models (Pearl 2009). Traditionally, influence diagrams have been used to help decision-makers make better decisions. Here, we invert the perspective, and use the diagrams to understand and predict the behaviour of machine

learning systems trained to optimize an objective in a given environment. To facilitate this analysis, we prove a number of relevant theorems and introduce two new concepts:

- *Value of Information* (VoI): First defined by Howard (1966), a graphical criterion for detecting positive VoI in influence diagrams were proposed and proven sound by Fagiuoli and Zaffalon (1998), Lauritzen and Nilsson (2001), and Shachter (2016). Here we offer the first correct completeness proof, showing that the graphical criterion is unique and cannot be further improved upon.

- *Value of Control* (VoC): Defined by Shachter (1986), Matheson (1990), and Shachter and Heckerman (2010), an incomplete graphical criterion was discussed by Shachter (1986). Here we provide a complete graphical criterion, along with both soundness and completeness proofs.

- *Instrumental Control incentive* (ICI): We propose a refinement of VoC to nodes the agent can influence with its decision. Conceptually, this is a hybrid of VoC and *responsiveness* (Shachter 2016). We offer a formal definition of instrumental control incentives based on nested counterfactuals, and establish a sound and complete graphical criterion.

- *Response incentive* (RI): Which changes in the environment does an optimal policy respond to? This is a central problem in fairness and AI safety (e.g. Kusner et al. 2017; Hadfield-Menell et al. 2017). Again, we give a formal definition, and a sound and complete graphical criterion.

Our analysis focuses on influence diagrams with a single-decision. This single-decision setting is adequate to model supervised learning, (contextual) bandits, and the choice of a policy in an MDP. Previous work has also discussed ways to transform a multi-decision setting into a single-decision setting by imputing policies to later decisions (Shachter 2016).

Applicability This paper combines material from two preprints (Everitt et al. 2019c; Carey et al. 2020). Since the release of these preprints, the unified language of causal influence diagrams have already aided in the understanding of incentive problems such as an agent's redirectability, ambition, tendency to tamper with reward, and other properties (Armstrong et al. 2020; Holtman 2020; Cohen, Vellambi, and Hutter 2020; Everitt et al. 2019a,b; Langlois and Everitt 2021).

(a) Fairness example: grade prediction   (b) Safety example: content recommendation
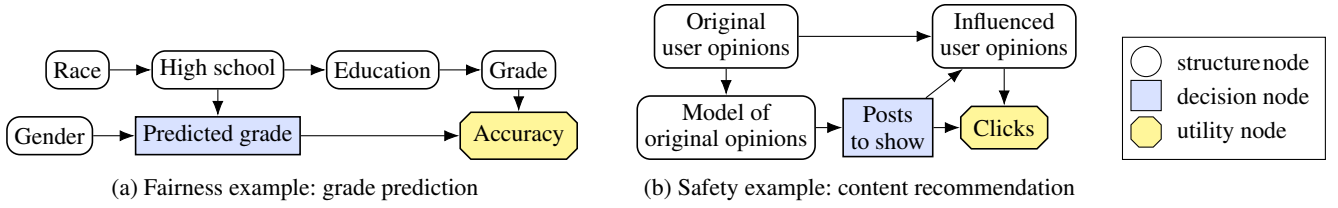
Figure 1: Two examples of decision problems represented as causal influence diagrams. In a) a predictor at a hypothetical university aims to estimate a student's grade, using as inputs their gender and the high school they attended. We ask whether the predictor is incentivised to behave in a discriminatory manner with respect to the students' gender and race. In this hypothetical cohort of students, performance is assumed to be a function of the quality of the high-school education they received. A student's high-school is assumed to be impacted by their race, and can affect the quality of their education. Gender, however, is assumed not to have an effect. In b) the goal of a content recommendation system is to choose posts that will maximise the user's click rate. However, the system's designers prefer the system not to manipulate the user's opinions in order to obtain more clicks.

Setup

To analyse agents' incentives, we will need a graphical framework with the causal properties of a structural causal model and the node categories of an influence diagram. This section will define such a model after reviewing structural causal models and influence diagrams.

Structural Causal Models

Structural causal models (SCMs) Pearl (2009) are a type of causal model where all randomness is consigned to *exogenous* variables, while deterministic *structural* functions relate the *endogenous* variables to each other and to the exogenous ones. As demonstrated by Pearl (2009), this *structural* approach has significant benefits over traditional causal Bayesian networks for analysing (nested) counterfactuals and "individual-level" effects.

Definition 1 (Structural causal model; Pearl 2009, Chapter 7). A *structural causal model* (with independent errors) is a tuple $\langle \mathcal{E}, V, F, P \rangle$, where $\mathcal{E}$ is a set of exogenous variables; V is a set of endogenous variables; and $F = \{f^V\}_{V \in V}$ is a collection of functions, one for each V. Each function $f^V \colon \operatorname{dom}(\mathbf{Pa}^V \cup \{\mathcal{E}^V\}) \to \operatorname{dom}(V)$ specifies the value of V in terms of the values of the corresponding exogenous variable $\mathcal{E}^V$ and endogenous parents $\mathbf{Pa}^V \subset V$, where these functional dependencies are acyclic. The domain of a variable V is $\operatorname{dom}(V)$ and for a set of variables, $\operatorname{dom}(W) := \times_{W \in W} \operatorname{dom}(W)$. The uncertainty is encoded through a probability distribution $P(\varepsilon)$ such that the exogenous variables are mutually independent.

For example, Figure 2b shows an SCM that models how *posts* (D) can influence a user's *opinion* (O) and *clicks* (U).x

The exogenous variables $\mathcal{E}$ of an SCM represent factors that are not modelled. For any value $\mathcal{E} = \varepsilon$ of the exogenous variables, the value of any set of variables $W \subseteq V$ is given by recursive application of the structural functions F and is denoted by $W(\varepsilon)$. Together with the distribution $P(\varepsilon)$ over exogenous variables, this induces a joint distribution $\Pr(W = w) = \sum_{\{\varepsilon | W(\varepsilon) = w\}} P(\varepsilon)$.

SCMs model *causal interventions* that set variables to particular values. These are defined via submodels:

Definition 2 (Submodel; Pearl 2009, Chapter 7). Let $\mathcal{M} = \langle \mathcal{E}, V, F, P \rangle$ be an SCM, X a set of variables in V, and x a particular realization of X. The submodel $\mathcal{M}_x$ represents the effects of an *intervention* $\operatorname{do}(X = x)$, and is formally defined as the SCM $\langle \mathcal{E}, V, F_x, P \rangle$ where $F_x = \{f^V | V \notin X\} \cup \{X = x\}$. That is to say, the original functional relationships of $X \in X$ are replaced with the constant functions $X = x$.

More generally, a *soft intervention* on a variable X in an SCM $\mathcal{M}$ replaces f^X with a function $g^X \colon \operatorname{dom}(\mathbf{Pa}^X \cup \{\mathcal{E}^X\}) \to \operatorname{dom}(X)$ (Eberhardt and Scheines 2007; Tian and Pearl 2001). The probability distribution $\Pr(W_{g^x})$ on any $W \subseteq V$ is defined as the value of $\Pr(W)$ in the submodel $\mathcal{M}_{g^x}$ where $\mathcal{M}_{g^x}$ is $\mathcal{M}$ modified by replacing f^X with g^X.

If W is a variable in an SCM $\mathcal{M}$, then W_x refers to the same variable in the submodel $\mathcal{M}_x$ and is called a *potential response variable*. In Figure 2b, the random variable O represents user opinion under "default" circumstances while O_d in Figure 2c represents the user's opinion given an intervention $\operatorname{do}(D = d)$ on the content posted. Note also how the intervention on D severs the link from ε^D to d in Figure 2c, as the intervention on D overrides the causal effect from D's parents. Throughout this paper we use subscripts to indicate submodels or interventions, and superscripts for indexing.

More elaborate hypotheticals can be described with a nested counterfactual, in which the intervention is itself a potential response variable. In Figure 2c, the *click* probability U depends on both the chosen *posts* D and the user *opinion* O, which is in turn also influenced by D. The nested potential response variable U_{O_d}, defined by $U_{O_d}(\varepsilon) := U_o(\varepsilon)$ where $o = O_d(\varepsilon)$, represents the probability that a user clicks on a "default" post D given that their opinion has been influenced by a hypothetical post d. In other words, the effect of the intervention $\operatorname{do}(D = d)$ is propagated to U only through O.

Causal Influence Diagrams

Influence diagrams are graphical models with special decision and utility nodes, developed to model decision making problems (Howard 1990; Lauritzen and Nilsson 2001). Influence diagrams do not in general have causal semantics, although some causal structure can be inferred (Heckerman and Shachter 1995). We will assume that the edges of the
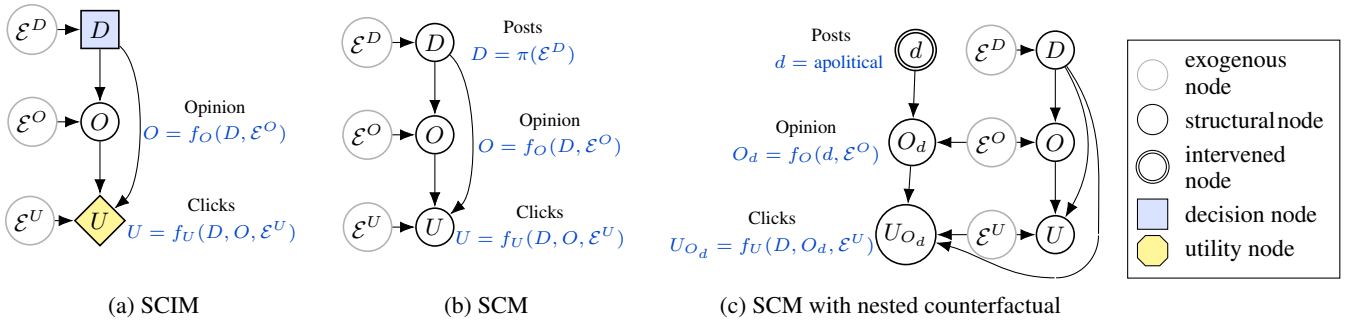
Figure 2: An example of a SCIM and interventions. In the SCIM, either political or apolitical posts D are displayed. These affect the user's opinion O. D and O influence the user's clicks U (a). Given a policy, the SCIM becomes a SCM (b). Interventions and counterfactuals may be defined in terms of this SCM. For example, the nested counterfactual U_{O_d} represents the number of clicks if the user has the opinions that they would arrive at, after viewing apolitical content (c).

influence diagram reflect the causal structure of the environment, so we use the term "Causal Influence Diagram".

Definition 3 (Causal influence diagram). A *causal influence diagram* (CID) is a directed acyclic graph $\mathcal{G}$ where the vertex set $\mathbf{V}$ is partitioned into *structure nodes* $\mathbf{X}$, *decision nodes* $\mathbf{D}$, and *utility nodes* $\mathbf{U}$. Utility nodes have no children.

We use $\mathbf{Pa}^V$ and $\mathbf{Desc}^V$ to denote the parents and descendants of a node $V \in \mathbf{V}$. The parents of the decision, $\mathbf{Pa}^D$, are also called observations. An edge from node V to node Y is denoted $V \to Y$. Edges into decisions are called information links, as they indicate what information is available at the time of the decision. A directed path (of length at least zero) is denoted $V \dashrightarrow Y$. For sets of variables, $\mathbf{V} \dashrightarrow \mathbf{Y}$ means that $V \dashrightarrow Y$ holds for some $V \in \mathbf{V}, Y \in \mathbf{Y}$.

Structural Causal Influence Models

For our new incentive concepts, we define a hybrid of the influence diagram and the SCM. Such a model, originally proposed by Dawid (2002), has structure and utility nodes with associated functions, exogenous variables with an associated probability distributions, and decision nodes, without any function at all, until one is selected by an agent.[1] This can be formalised as the *structural causal influence model* (SCIM, pronounced 'skim').

Definition 4 (Structural causal influence model). A *structural causal influence model* (SCIM) is a tuple $\mathcal{M} = \langle \mathcal{G}, \boldsymbol{\mathcal{E}}, \boldsymbol{F}, P \rangle$ where:

- $\mathcal{G}$ is a CID with finite-domain variables $\mathbf{V}$ (partitioned into $\mathbf{X}, \mathbf{D}$, and $\mathbf{U}$) where utility variable domains are a subset of $\mathbb{R}$. We say that $\mathcal{M}$ is *compatible with* $\mathcal{G}$.
- $\boldsymbol{\mathcal{E}} = \{\mathcal{E}^V\}_{V \in \mathbf{V}}$ is a set of finite-domain *exogenous variables*, one for each endogenous variable.
- $\boldsymbol{F} = \{f^V\}_{V \in \mathbf{V} \setminus \mathbf{D}}$ is a set of *structural functions* $f^V \colon \mathrm{dom}(\mathbf{Pa}^V \cup \{\mathcal{E}^V\}) \to \mathrm{dom}(V)$ that specify how each non-decision endogenous variable depends on its parents in $\mathcal{G}$ and its associated exogenous variable.

- P is a probability distribution for $\boldsymbol{\mathcal{E}}$ such that the individual exogenous variables $\mathcal{E}^V$ are mutually independent.

We will restrict our attention to single-decision settings with $\mathbf{D} = \{D\}$. An example of such a SCIM for the content recommendation example is shown in Figure 2a. In single-decision SCIMs, the decision-making task is to maximize expected utility by selecting a decision $d \in \mathrm{dom}(D)$ based on the observations $\mathbf{Pa}^D$. More formally, the task is to select a structural function for D in the form of a *policy* $\pi \colon \mathrm{dom}(\mathbf{Pa}^D \cup \{\mathcal{E}^D\}) \to \mathrm{dom}(D)$. The exogenous variable $\mathcal{E}^D$ provides randomness to allow the policy to be a stochastic function of its endogenous parents $\mathbf{Pa}^D$. The specification of a policy turns a SCIM $\mathcal{M}$ into an SCM $\mathcal{M}_\pi := \langle \boldsymbol{\mathcal{E}}, \mathbf{V}, \boldsymbol{F} \cup \{\pi\}, P \rangle$, see Figure 2b. With the resulting SCM, the standard definitions of causal interventions apply. Note that what determines whether a node is observed or not at the time of decision-making is whether the node is a parent of the decision. Commonly, some structure nodes represent latent variables that are unobserved.

We use Pr_π and $\mathbb{E}_\pi$ to denote probabilities and expectations with respect to $\mathcal{M}_\pi$. For a set of variables $\mathbf{X}$ not in $\mathbf{Desc}^D$, $\mathrm{Pr}_\pi(\boldsymbol{x})$ is independent of π and we simply write $\mathrm{Pr}(\boldsymbol{x})$. An *optimal policy* for a SCIM is defined as any policy π that maximises $\mathbb{E}_\pi[\mathcal{U}]$, where $\mathcal{U} := \sum_{U \in \mathbf{U}} U$. A potential response $\mathcal{U}_{\boldsymbol{x}}$ is defined as $\mathcal{U}_{\boldsymbol{x}} := \sum_{U \in \mathbf{U}} U_{\boldsymbol{x}}$.

Materiality

Next, we review a characterization of which observations are *material* for optimal performance, as this will be a fundamental building block for most of our theory.[2]

Definition 5 (Materiality; Shachter 2016). For any given SCIM $\mathcal{M}$, let $\mathcal{V}^*(\mathcal{M}) = \max_\pi \mathbb{E}_\pi[\mathcal{U}]$ be the maximum attainable utility in $\mathcal{M}$, and let $\mathcal{M}_{X \not\to D}$ be $\mathcal{M}$ modified by removing any information link $X \to D$. The observation $X \in \mathbf{Pa}^D$ is *material* if $\mathcal{V}^*(\mathcal{M}_{X \not\to D}) < \mathcal{V}^*(\mathcal{M})$.

Nodes may often be identified as immaterial based on the graphical structure alone (Fagiuoli and Zaffalon 1998;

---

[1]Dawid called this a "functional influence diagram". We favour the term SCIM, because the corresponding term SCM is more prevalent than "functional model".

[2]In contrast to subsequent sections, the results in this section and the VoI section do not require the influence diagrams to be causal.

Lauritzen and Nilsson 2001; Shachter 2016). The graphical criterion uses uses the notion of d-separation.

Definition 6 (d-separation; Verma and Pearl 1988)**.** A path p is said to be d-separated by a set of nodes $\mathbf{Z}$ if and only if:

1. p contains a collider $X \rightarrow W \leftarrow Y$, such that the middle node W is not in $\mathbf{Z}$ and no descendants of W are in $\mathbf{Z}$, or

2. p contains a chain $X \rightarrow W \rightarrow Y$ or fork $X \leftarrow W \rightarrow Y$ where W is in $\mathbf{Z}$, or

3. one or both of the endpoints of p is in $\mathbf{Z}$.

A set $\mathbf{Z}$ is said to d-separate $\mathbf{X}$ from $\mathbf{Y}$, written $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if $\mathbf{Z}$ d-separates every path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$. Sets that are not d-separated are called d-connected.

According to the graphical criterion of Fagiuoli and Zaffalon (1998), an observation cannot provide useful information if it is d-separated from utility, conditional on other observations. This condition is called nonrequisiteness.

Definition 7 (Nonrequisite observation; Lauritzen and Nilsson 2001)**.** Let $\mathbf{U}^D := \mathbf{U} \cap \mathbf{Desc}^D$ be the utility nodes downstream of D. An observation $X \in \mathbf{Pa}^D$ in a single-decision CID $\mathcal{G}$ is *nonrequisite* if:

$$X \perp \mathbf{U}^D \mid \left(\mathbf{Pa}^D \cup \{D\} \setminus \{X\} \right) \tag{1}$$

In this case, the edge $X \rightarrow D$ is also called nonrequisite. Otherwise X and $X \rightarrow D$ are *requisite*.

For example, in Figure 3a, *high school* is a requisite observation while *gender* is not.

Value of Information

Materiality can be generalized to nodes not observed, to assess which variables a decision-maker would benefit from knowing before making a decision, i.e. which variables have VoI (Howard 1966; Matheson 1990). To assess VoI for a variable X, we first make X an observation by adding a link $X \rightarrow D$, and then test whether X is material in the updated model (Shachter 2016).

Definition 8 (Value of information)**.** A node $X \in \mathbf{V} \setminus \mathbf{Desc}^D$ in a single-decision SCIM $\mathcal{M}$ has *VoI* if it is material in the model $\mathcal{M}_{X \rightarrow D}$ obtained by adding the edge $X \rightarrow D$ to $\mathcal{M}$. A CID $\mathcal{G}$ *admits VoI* for X if X has VoI in a a SCIM $\mathcal{M}$ compatible with $\mathcal{G}$.

Since Definition 8 adds an information link, it can only be applied to non-descendants of the decision, lest cycles be created in the graph. Fortunately, the structural functions need not be adapted for the added link, since there is no structural function associated with D.

We prove that the graphical criterion of Definition 7 is tight for both materiality and VoI, in that it identifies every zero VoI node that can be identified from the graphical structure (in a single decision setting).

Theorem 9 (Value of information criterion)**.** *A single decision CID $\mathcal{G}$ admits VoI for $X \in \mathbf{V} \setminus \mathbf{Desc}^D$ if and only if X is a requisite observation in $\mathcal{G}_{X \rightarrow D}$, the graph obtained by adding $X \rightarrow D$ to $\mathcal{G}$.*

The soundness direction (i.e. the *only if* direction) follows from d-separation (Fagiuoli and Zaffalon 1998; Lauritzen and Nilsson 2001; Shachter 2016). In contrast, the completeness direction does not follow from the completeness property of d-separation. The d-connectedness of X to $\mathbf{U}$ implies that $\mathbf{U}$ may be conditionally dependent on X. It does not imply, however, that the expectation of $\mathbf{U}$ or the utility attainable under an optimal policy will change. Instead, our proof in the supplementary material[3] constructs a SCIM such that X is material. This differs from a previous attempt by Nielsen and Jensen (1999), as discussed in Related Work.

We apply the graphical criterion to the grade prediction example in Figure 3a. One can see that the predictor has an incentive to use the incoming student's high school but not gender. This makes intuitive sense, given that *gender* provides no information useful for predicting the university grade in this example.

Response Incentives

There are two ways to understand a material observation. One is that it provides useful information. From this perspective, a natural generalisation is VoI, as described in the previous section. An alternative perspective is that a material observation is one that influences optimal decisions. Under this interpretation, the natural generalisation is the set of all (observed and unobserved) variables that influence the decision. We say that these variables have a response incentive.[4]

Definition 10 (Response incentive)**.** Let $\mathcal{M}$ be a single-decision SCIM. A policy π *responds* to a variable $X \in \mathbf{X}$ if there exists some intervention $\mathrm{do}(X = x)$ and some setting $\boldsymbol{\mathcal{E}} = \boldsymbol{\varepsilon}$, such that $D_x(\boldsymbol{\varepsilon}) \neq D(\boldsymbol{\varepsilon})$. The variable X has a *response incentive* if all optimal policies respond to X.

A CID *admits* a response incentive on X if it is compatible with a SCIM that has a response incentive on X.

For a response incentive on X to be possible, there must be: i) a directed path $X \dashrightarrow D$, and ii) an incentive for D to use information from that path. For example, in Figure 3a, *gender* has a directed path to the decision but it does not provide any information about the likely grade, so there is no response incentive. The graphical criterion for RI builds on a modified graph with nonrequisite information links removed.

Definition 11 (Minimal reduction; Lauritzen and Nilsson 2001)**.** The *minimal reduction* $\mathcal{G}^{\min}$ of a single-decision CID $\mathcal{G}$ is the result of removing from $\mathcal{G}$ all information links from nonrequisite observations.

The presence (or absence) of a path $X \dashrightarrow D$ in the minimal reduction tells us whether a response incentive can occur.

Theorem 12 (Response incentive criterion)**.** *A single-decision CID $\mathcal{G}$ admits a response incentive on $X \in \mathbf{X}$ if and only if the minimal reduction $\mathcal{G}^{min}$ has a directed path $X \dashrightarrow D$.*

---

[3]Available at https://arxiv.org/abs/2102.01685

[4]The term *responsiveness* (Heckerman and Shachter 1995; Shachter 2016) has a related but not identical meaning – it refers to whether a decision D affects a variable X rather than whether X affects D.

(a) Admits response incentive on race

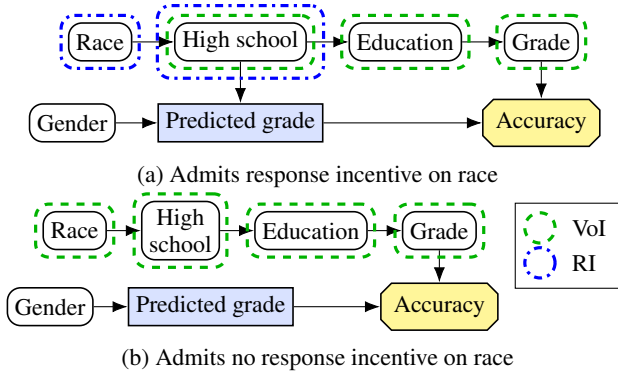(b) Admits no response incentive on race

Figure 3: In (a), the admissible incentives of the grade prediction example from Figure 1a are shown, including a response incentive on race. In (b), the predictor no-longer has access to the students' high school, and hence there can no-longer be any response incentive on race.

Proof. The *if* (completeness) direction is proved in the supplementary material. For the soundness direction, assume that for $\mathcal{G}$, the minimal reduction $\mathcal{G}^{\min}$ does not contain a directed path $X \dashrightarrow D$. Let $\mathcal{M} = \langle \mathcal{G}, \mathcal{E}, \mathbf{F}, P \rangle$ be any SCIM compatible with $\mathcal{G}$. Let $\mathcal{M}^{\min} = \langle \mathcal{G}^{\min}, \mathcal{E}, \mathbf{F}, P \rangle$ be $\mathcal{M}$, but with the minimal reduction $\mathcal{G}^{\min}$. By Lemma 25 in the supplementary material, there exists a $\mathcal{G}^{\min}$-respecting policy $\tilde{\pi}$ that is optimal in $\mathcal{M}$. In $\mathcal{M}^{\min}_{\tilde{\pi}}$, X is causally irrelevant for D so $D(\varepsilon) = D_x(\varepsilon)$. Furthermore, $\mathcal{M}_{\tilde{\pi}}$ and $\mathcal{M}^{\min}_{\tilde{\pi}}$ are the same SCM, with the functions $\mathbf{F} \cup \{\tilde{\pi}\}$. So $D(\varepsilon) = D_x(\varepsilon)$ also in $\mathcal{M}_{\tilde{\pi}}$, which means that there is an optimal policy in $\mathcal{M}$ that does not respond to interventions on X for any ε. $\square$

The intuition behind the proof is that an optimal decision only responds to effects that propagate to one of its requisite observations. For the completeness direction, we show in the supplementary material that if $X \dashrightarrow D$ is present in the minimal reduction $\mathcal{G}^{\min}$, then we can select a SCIM $\mathcal{M}$ compatible with $\mathcal{G}$ such that D receives useful information along that path, that any optimal policy must respond to.

In a safety setting, it may be desirable for an AI system to have an incentive to respond to its shutdown button, so that when asked to shut down, it does so (Hadfield-Menell et al. 2017). In a fairness setting, on the other hand, a response incentive may be a cause for concern, as illustrated next.

Incentivised unfairness   Response incentives are closely related to counterfactual fairness (Kusner et al. 2017; Kilbertus et al. 2017). A prediction — or more generally a decision — is considered counterfactually unfair if a change to a *sensitive attribute* like race or gender would change the decision.

Definition 13 (Counterfactual fairness; Kusner et al. 2017)**.** A policy π is *counterfactually fair* with respect to a sensitive attribute A if

$$\Pr_{\pi}\left(D_{a'} = d \mid \mathbf{pa}^D, a\right) = \Pr_{\pi}\left(D = d \mid \mathbf{pa}^D, a\right)$$

for every decision $d \in \mathrm{dom}(D)$, every context $\mathbf{pa}^D \in \mathrm{dom}(\mathbf{Pa}^D)$, and every pair of attributes $a, a' \in \mathrm{dom}(A)$ with

$\Pr(\mathbf{pa}^D, a) > 0$.

A response incentive on a sensitive attribute indicates that counterfactual unfairness is incentivised, as it implies that *all* optimal policies are counterfactually unfair:

Theorem 14 (Counterfactual fairness and response incentives)**.** *In a single-decision SCIM $\mathcal{M}$ with a sensitive attribute $A \in \mathbf{X}$, all optimal policies π^* are counterfactually unfair with respect to A if and only if A has a response incentive.*

The proof is given in the supplementary material.

A response incentive on a sensitive attribute means that counterfactual unfairness is not just possible, but incentivised. As a result, it has a more restrictive graphical criterion. The graphical criterion for counterfactual fairness states that a decision can only be counterfactually unfair with respect to a sensitive attribute if that attribute is an ancestor of the decision (Kusner et al. 2017, Lemma 1). For example, in the grade prediction example of Figure 3a, it is possible for a predictor to be counterfactually unfair with respect to either *gender* or *race*, because both are ancestors of the decision. The response incentive criterion can tell us in which case counterfactual unfairness is actually incentivised. In this example, the minimal reduction includes the edge from *high school* to *predicted grade* and hence the directed path from *race* to *predicted grade*. However, it excludes the edge from *gender* to *predicted grade*. This means that the agent is incentivised to be counterfactually unfair with respect to *race* but not to *gender*.

Based on this, how should the system be redesigned? According to the response incentive criterion, the most important change is to remove the path from *race* to *predicted grade* in the minimal reduction. This can be done by removing the agent's access to *high school*. This change is implemented in Figure 3b, where there is no response incentive on either sensitive variable.

Value of information is also related to fairness. For a sensitive variable that is not a parent of the decision, positive VoI means that *if* the predictor gained access to its value, then the predictor would use it. For example, if in Figure 3b an edge is added from *race* to *predicted grade*, then unfair behaviour will result. In practice, such access can result from unanticipated correlations between the sensitive attribute and parents of the decision, rather than the system being given direct access to the attribute. Analysing VoI may help detect such problems at an early stage. However, VoI is less closely related to counterfactual fairness than response incentives. In particular, *race* lacks VoI in Figure 3a, but counterfactual unfairness is incentivised. On the other hand, Figure 3b admits positive VoI for *race*, but counterfactual unfairness is not incentivised.

The incentive approach is not restricted to counterfactual fairness. For any fairness definition, one could assess whether that kind of unfairness is incentivised by checking whether it is present under all optimal policies.

Value of Control

A variable has VoC if a decision-maker could benefit from setting its value (Shachter 1986; Matheson 1990; Shachter

and Heckerman 2010). Concretely, we ask whether the attainable utility can be increased by letting the agent decide the structural function for the variable.

Definition 15 (Value of control). In a single-decision SCIM $\mathcal{M}$, a non-decision node X has *positive value of control* if

$$\max_\pi \mathbb{E}_\pi[\mathcal{U}] < \max_{\pi, g^X} \mathbb{E}_\pi[\mathcal{U}_{g^X}]$$

where $g^X : \mathrm{dom}(\mathbf{Pa}^X \cup \{\mathcal{E}^X\}) \to \mathrm{dom}(X)$ is a soft intervention at X, i.e. a new structural function for X that respects the graph.

A CID $\mathcal{G}$ admits *positive value of control* for X if there exists a SCIM $\mathcal{M}$ compatible with $\mathcal{G}$ where X has positive value of control. This can be deduced from the graph, using again the minimal reduction (Definition 11) to rule out effects through observations that an optimal policy can ignore.

Theorem 16 (Value of control criterion). *A single-decision CID $\mathcal{G}$ admits positive value of control for a node $X \in \mathbf{V} \setminus \{D\}$ if and only if there is a directed path $X \dashrightarrow \mathbf{U}$ in the minimal reduction $\mathcal{G}^{min}$.*

Proof. The *if* (completeness) direction is proved in the supplementary material. The proof of *only if* (soundness) is as follows. Let $\mathcal{M} = \langle \mathcal{G}, \mathcal{E}, \mathbf{F}, P \rangle$ be a single-decision SCIM. Let $\mathcal{M}_{g^X}$ be $\mathcal{M}$, but with the structural function f^X replaced with g^X. Let $\mathcal{M}^{min}$ and $\mathcal{M}_{g^X}^{min}$ be the same SCIMs, respectively, but replacing each graph with the minimal reduction $\mathcal{G}^{min}$.

Recall that $\mathbb{E}_\pi[\mathcal{U}_{g^X}]$ is defined by applying the soft intervention g^X to the (policy-completed) SCM $\mathcal{M}_\pi$. However, this is equivalent to applying the policy π to the modified SCIM $\mathcal{M}_{g^X}$, as the resulting SCMs are identical. Since $\mathcal{M}_{g^X}$ is a SCIM, Lemma 25 in the supplementary material can be applied, to find a $\mathcal{G}^{min}$-respecting optimal policy $\tilde{\pi}$ for $\mathcal{M}_{g^X}$.

Consider now the expected utility under an arbitrary intervention g^X for a policy π optimal for $\mathcal{M}_{g^X}$:

$$
\begin{aligned}
&\mathbb{E}_\pi[\mathcal{U}_{g^X}] \text{ in } \mathcal{M} \\
&= \mathbb{E}_\pi[\mathcal{U}] \text{ in } \mathcal{M}_{g^X} && \text{by SCM equivalence} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}_{g^X} && \text{by Lemma 25} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}_{g^X}^{min} && \text{since } \tilde{\pi} \text{ is } \mathcal{G}^{min}\text{-respecting} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}^{min} && \text{by Lemma 23} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M} && \text{only increasing the policy set} \\
&\leq \max_{\pi^*} \mathbb{E}_{\pi^*}[\mathcal{U}] \text{ in } \mathcal{M} && \text{max dominates all elements.}
\end{aligned}
$$

This shows that X must lack value of control. $\square$

The proof of the completeness direction in the supplementary material establishes that if a path exists, then a SCIM be selected where the intervention on X can either directly control U or increase the useful information available at D.

To apply this criterion to the content recommendation example (Figure 4a), we first obtain the minimal reduction, which is identical to the original graph. Since all non-decision nodes are upstream of the utility in the minimal reduction, they all admit positive VoC. Notably, this includes nodes like

original user opinions and *model of user opinions* that the decision has no ability to control according to the graphical structure. In the next section, we propose *instrumental control incentives*, which incorporate the agent's limitations.

Instrumental Control Incentive

Would an agent use its decision to control a variable X? This question has two parts: whether X is useful to control (VoC), and whether X is possible to control (responsiveness). As described in the previous section, VoC uses $\mathcal{U}_{g^X}$ to consider the utility attainable from arbitrary control of X. Meanwhile, X_d describes the way X can be controlled by D. These notions can be combined with a nested counterfactual $\mathcal{U}_{X_d}$, which expresses the effect that D can have on $\mathcal{U}$ by controlling X.

Definition 17 (Instrumental control incentive). In a single-decision SCIM $\mathcal{M}$, there is an *instrumental control incentive* on a variable X in decision context $\mathbf{pa}^D$ if, for all optimal policies π^*,

$$\mathbb{E}_{\pi^*}[\mathcal{U}_{X_d} \mid \mathbf{pa}^D] \neq \mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{pa}^D]. \tag{2}$$

Conceptually, an instrumental control incentive can be interpreted as follows. If the agent got to choose D to influence X independently of how D influences other aspects of the environment, would that choice matter? We call it an *instrumental* control incentive, as the control of X is a tool for achieving utility (cf. *instrumental goals* Omohundro 2008; Bostrom 2014). ICIs do not consider side-effects of the optimal policy: for instance, it may be that all optimal policies affect X in a particular way, even if X is a not an ancestor of any utility node — in such cases, no ICI is present. Finally, in Pearl's (2001) terminology, an instrumental control incentive corresponds to a *natural indirect effect* from D to U via X in $\mathcal{M}_{\pi^*}$, for all optimal policies π^*.

A CID $\mathcal{G}$ *admits* an instrumental control incentive on X if $\mathcal{G}$ is compatible with a SCIM $\mathcal{M}$ with an instrumental control incentive on X for some decision context $\mathbf{pa}^D$. The following theorem gives a sound and complete graphical criterion for which CIDs admit instrumental control incentives.

Theorem 18 (Instrumental Control Incentive Criterion). *A single-decision CID $\mathcal{G}$ admits an instrumental control incentive on $X \in \mathbf{V}$ if and only if $\mathcal{G}$ has a directed path from the decision D to a utility node $U \in \mathbf{U}$ that passes through X, i.e. a directed path $D \dashrightarrow X \dashrightarrow \mathbf{U}$.*

Proof. Completeness (the *if* direction) is proved in the supplementary material. The proof of soundness is as follows.

Let $\mathcal{M}$ be any SCIM compatible with $\mathcal{G}$ and π any policy for $\mathcal{M}$. We consider variables in the SCM $\mathcal{M}_\pi$. If there is no directed path $D \dashrightarrow X \dashrightarrow \mathbf{U}$ in $\mathcal{G}$, then either $D \not\dashrightarrow X$ or $X \not\dashrightarrow \mathbf{U}$. If $D \not\dashrightarrow X$, then $X_d(\varepsilon) = X(\varepsilon)$ for any setting $\varepsilon \in \mathrm{dom}(\mathcal{E})$ and decision d (Lemma 20 in the supplementary material). Therefore, $\mathcal{U}(\varepsilon) = \mathcal{U}_{X_d}(\varepsilon)$. Similarly, if $X \not\dashrightarrow \mathbf{U}$ then $U(\varepsilon) = U_x(\varepsilon)$ for every setting $\varepsilon \in \mathrm{dom}(\mathcal{E})$, $x \in \mathrm{dom}(X)$ and $U \in \mathbf{U}$ so $\mathcal{U}(\varepsilon) = \mathcal{U}_{X_d}(\varepsilon)$. In either case, $\mathbb{E}_\pi[\mathcal{U} \mid \mathbf{pa}^D] = \mathbb{E}_\pi[\mathcal{U}_{X_d} \mid \mathbf{pa}^D]$ and there is no instrumental control incentive on X. $\square$

The logic behind the soundness proof above is that if there is no path from D to X to $\mathbf{U}$, then D cannot have any effect

(a) Admits instrumental control incentive on user opinion

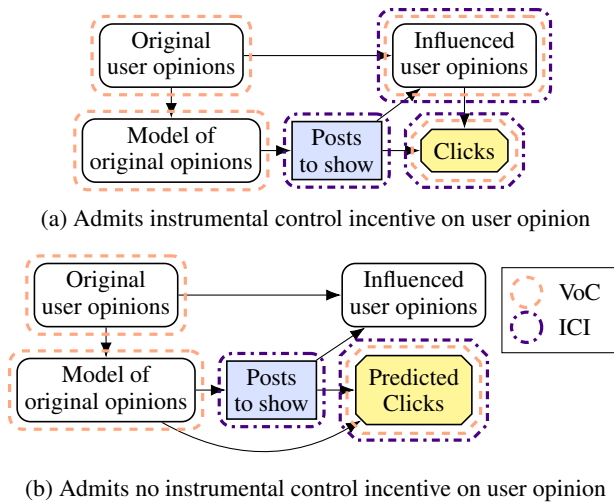(b) Admits no instrumental control incentive on user opinion

Figure 4: In (a), the content recommendation example from Figure 1b is shown to admit an instrumental control incentive on user opinion. This is avoided in (b) with a change to the objective.

on U via X. For the completeness direction proved in the supplementary material, we show how to construct a SCIM so that U_{X_d} differs from the non-intervened U for any diagram with a path $D \dashrightarrow X \dashrightarrow U$.

Let us apply this criterion to the content recommendation example in Figure 4a. The only nodes X in this graph that lie on a path $D \dashrightarrow X \dashrightarrow U$ are *clicks* and *influenced user opinions*. Since *influenced user opinions* has an instrumental control incentive, the agent may seek to influence that variable in order to attain utility. For example, it may be easier to predict what content a more emotional user will click on and therefore, a recommender may achieve a higher click rate by introducing posts that induce strong emotions.

How could we instead design the agent to maximise clicks without manipulating the user's opinions (i.e. without an instrumental control incentive on *influenced user opinions*)? As shown in Figure 4b, we could redesign the system so that instead of being rewarded for the true click rate, it is rewarded for the clicks it would be predicted to have, based on a separately trained model of the user's preferences. An agent trained in this way would view any modification of user opinions as irrelevant for improving its performance; however, it would still have an instrumental control incentive for *predicted clicks* so it would still deliver desired content. To avoid undesirable behaviour in practice, the click prediction must truly predict whether the original user would click the content, rather than baking in the effect of changes to the user's opinion from reading earlier posts. This could be accomplished, for instance, by training a model to predict how many clicks each post would receive if it was offered individually.

This dynamic is related to concerns about the long-term safety of AI systems. For example, Russell (2019) has hypothesised that an advanced AI system would seek to manipulate its objective function (or human overseer) to obtain reward.

This can be understood as an instrumental control incentive on the objective function (or the overseer's behaviour). A better understanding of incentives could therefore be relevant for designing safe systems in both the short and long-term.

Related Work

Causal influence diagrams Jern and Kemp (2011) and Kleiman-Weiner et al. (2015) define influence diagrams with causal edges, and similarly use them to model decision-making of rational agents (although they are less formal than us, and focus on human decision-making).

An informal precursor of the SCIM that also used structural functions (as opposed to conditional probability distributions) was the "functional influence diagram" (Dawid 2002). The most similar alternative model is the Howard canonical form influence diagram (Howard 1990; Heckerman and Shachter 1995). However, this only permits counterfactual reasoning downstream of decisions, which is inadequate for defining the response incentive. Similarly, the causality property for influence diagrams introduced by Heckerman and Shachter (1994) and Shachter and Heckerman (2010) only constrains the relationships to be partially causal downstream of the decision (though adding new decision-node parents to all nodes makes the diagram fully causal). Appendix A in the supplementary material shows by example why the stronger causality property is necessary for most of our incentive concepts.

An open-source Python implementation of CIDs has recently been developed[5] (Fox et al. 2021).

Value of information and control Theorems 9 and 16 for value of information and value of control build on previous work. The concepts were first introduced by Howard (1966) and Shachter (1986), respectively. The VoI soundness proof follows previous proofs (Shachter 1998; Lauritzen and Nilsson 2001), while the VoI completeness proof is most similar to an attempted proof by Nielsen and Jensen (1999). They propose the criterion $X \not\perp U^D \mid \mathbf{Pa}_D$ for requisite nodes, which differs from (1) in the conditioned set. Taken literally,[6] their criterion is unsound for requisite nodes and positive VoI. For example, in Figure 3a, *High school* is d-separated from *accuracy* given $\mathbf{Pa}^D$, so their criterion would fail to detect that *High school* is requisite and admits VoI.[7][8]

---

[5]https://github.com/causalincentives/pycid

[6]Def. 6 defines d-separation for potentially overlapping sets.

[7]Furthermore, to prove that nodes meeting the d-connectedness property are requisite, Nielsen and Jensen claim that "X is [requisite] for D if $\Pr(\mathrm{dom}(U) \mid D, \mathbf{Pa}^D)$ is a function of X and U is a utility function relevant for D". However, U being a function of X only proves that U is conditionally dependent on X, not that it changes the expected utility, or is requisite or material. Additional argumentation is needed to show that conditioning on X can actually change the expected utility; our proof provides such an argument.

[8]Since a preprint of this paper was placed online (Everitt et al. 2019c), this completeness result was independently discovered by Zhang, Kumor, and Bareinboim (2020, Thm. 2) and Lee and Bareinboim (2020, Thm. 1). Theorem 2 in the latter also provides a crite-

To have positive VoC, it is known that a node must be an ancestor of a value node (Shachter 1986), but the authors know of no more-specific criterion. The concept of a *relevant* node introduced by Nielsen and Jensen (1999) also bears some semblance to VoC.

The relation of the current technical results to prior work is summarised in Table S1 in the Appendix.

Instrumental control incentives Kleiman-Weiner et al. (2015) use (causal) influence diagrams to define a notion of *intention*, that captures which nodes an optimal policy seeks to influence. Intention is conceptually similar to instrumental control incentives and uses hypothetical node deletions to ask which nodes the agent intends to control. Their concept is more refined than ICI in the sense that it includes includes only the nodes that determine optimal policy behaviour, but the definition is not properly formalized and it is not clear that it can be applied to all influence diagram structures.

AI fairness Another application of this work is to evaluate when an AI system is incentivised to behave unfairly, on some definition of fairness. Response incentives address this question for counterfactual fairness (Kusner et al. 2017; Kilbertus et al. 2017). An incentive criterion corresponding to path-specific effects (Zhang, Wu, and Wu 2017; Nabi and Shpitser 2018) is deferred to future work. Nabi, Malinsky, and Shpitser (2019) have shown how a policy may be chosen subject to path-specific effect constraints. However, they assume recall of all past events, whereas the response incentive criterion applies to any CID.

Mechanism design The aim of mechanism design is to understand how objectives and environments can be designed, in order to shape the behavior of rational agents (e.g. Nisan et al. 2007, Part II). At this high level, mechanism design is closely related to the incentive design results we have developed in this paper. In practice, the strands of research look rather different. The core challenge of mechanism design is that agents have private information or preferences. As we take the perspective of an agent designer, private information is only relevant for us to the extent that some types of agents or objectives may be harder to implement than others. Instead, our core challenge comes from causal relationships in agent environments, a consideration of little interest to most of mechanism design.

Discussion and Conclusion

We have proved sound and complete graphical criteria for two existing concepts (VoI and VoC) and two new concepts: response incentive and instrumental control incentive. The results have all focused on the (causal) structure of the interaction between agent and environment. This is both a strength and a weakness. On the one hand, it means that formal conclusions can be made about a system's incentives, even when details about the quantitative relationship between variables is unknown. On the other hand, it also means that

---

rion for material observations in a multi-decision setting.

these results will not help with subtler comparisons, such as the relative strength of different incentives. It also means that the causal relationships between variables must be known. This challenge is common to causal models in general. In the context of incentive design, it is partially alleviated by the fact that causal relationships often follow directly from the design choices for an agent and its objective. Finally, causal diagrams struggle to express dynamically changing causal relationships.

While important to be aware of, these limitations do not prevent causal influence diagrams from providing a clear, useful, and unified perspective on agent incentives. It has seen applications ranging from value learning (Armstrong et al. 2020; Holtman 2020), interruptibility (Langlois and Everitt 2021), conservatism (Cohen, Vellambi, and Hutter 2020), modeling of agent frameworks (Everitt et al. 2019b), and reward tampering (Everitt et al. 2019a). Through such applications, we hope that the incentive analysis described in this paper will ultimately contribute to more fair and safe AI systems.

References

Armstrong, S.; Orseau, L.; Leike, J.; and Legg, S. 2020. Pitfalls in learning a reward function online. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Carey, R.; Langlois, E.; Everitt, T.; and Legg, S. 2020. The Incentives that Shape Behaviour. In *SafeAI AAAI workshop*.

Cohen, M. K.; Vellambi, B. N.; and Hutter, M. 2020. Asymptotically Unambitious Artificial General Intelligence. In *AAAI Conference on Artificial Intelligence*.

Dawid, A. P. 2002. Influence diagrams for causal modelling and inference. *International Statistical Review* .

Eberhardt, F.; and Scheines, R. 2007. Interventions and causal inference. *Philosophy of Science* .

Everitt, T.; Hutter, M.; Kumar, R.; and Krakovna, V. 2019a. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *CoRR* .

Everitt, T.; Kumar, R.; Krakovna, V.; and Legg, S. 2019b. Modeling AGI Safety Frameworks with Causal Influence Diagrams. In *Workshop on Artificial Intelligence Safety*, volume 2419 of *CEUR Workshop Proceedings*.

Everitt, T.; Ortega, P. A.; Barnes, E.; and Legg, S. 2019c. Understanding Agent Incentives using Causal Influence Diagrams. Part I: Single Action Settings. *CoRR* .

Fagiuoli, E.; and Zaffalon, M. 1998. A note about redundancy in influence diagrams. *International Journal of Approximate Reasoning* .

Fox, J.; Hammond, L.; Everitt, T.; Abate, A.; and Wooldridge, M. 2021. Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice. In *AAMAS*.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. J. 2017. The Off-Switch Game. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Heckerman, D.; and Shachter, R. 1994. A Decision-Based View of Causality. In *Uncertainty in Artificial Intelligence (UAI)*, 302–310.

Heckerman, D.; and Shachter, R. D. 1995. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research* 3: 405–430. doi:10.1613/jair.202.

Holtman, K. 2020. AGI Agent Safety by Iteratively Improving the Utility Function. *International Conference on Artificial General Intelligence* .

Howard, R. A. 1966. Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics* .

Howard, R. A. 1990. From influence to relevance to knowledge. *Influence diagrams, belief nets and decision analysis* .

Jern, A.; and Kemp, C. 2011. Capturing mental state reasoning with influence diagrams. In *Proceedings of the 2011 Cognitive Science Conference*, 2498–2503.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

Kleiman-Weiner, M.; Gerstenberg, T.; Levine, S.; and Tenenbaum, J. B. 2015. Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 1123–1128.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*.

Langlois, E.; and Everitt, T. 2021. How RL Agents Behave when their Actions are Modified. In *AAAI*.

Lauritzen, S. L.; and Nilsson, D. 2001. Representing and Solving Decision Problems with Limited Information. *Management Science* .

Lee, S.; and Bareinboim, E. 2020. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in Neural Information Processing Systems* 33.

Matheson, J. E. 1990. Using influence diagrams to value information and control. In Oliver, R. M.; and Smith, J. Q., eds., *Influence Diagrams, Belief Nets, and Decision Analysis*. Wiley and Sons.

Nabi, R.; Malinsky, D.; and Shpitser, I. 2019. Learning optimal fair policies. *Proceedings of machine learning research* .

Nabi, R.; and Shpitser, I. 2018. Fair Inference on Outcomes. In *AAAI Conference on Artificial Intelligence*.

Nielsen, T. D.; and Jensen, F. V. 1999. Welldefined Decision Scenarios. In *Uncertainty in Artificial Intelligence (UAI)*.

Nisan, N.; Roughgarden, T.; Tardos, E.; and Vijay V Vazirani, eds. 2007. *Algorithmic Game Theory*. Cambridge University Press.

Omohundro, S. M. 2008. The Basic AI Drives. In Wang, P.; Goertzel, B.; and Franklin, S., eds., *Artificial General Intelligence*, volume 171. IOS Press.

O'Neil, C. 2016. *Weapons of Math Destruction*. Crown Books.

Pearl, J. 2001. Direct and Indirect Effects. In *Uncertainty in Artificial Intelligence (UAI)*.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition edition. ISBN 9780521895606.

Russell, S. J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Shachter, R. D. 1986. Evaluating Influence Diagrams. *Operations Research* .

Shachter, R. D. 1998. Bayes-Ball: The Rational Pastime (for Determining Irrelevance and Requisite Information in Belief Networks and Influence Diagrams). *Uncertainty in Artificial Intelligence (UAI)* .

Shachter, R. D. 2016. Decisions and Dependence in Influence Diagrams. In *International Conference on Probabilistic Graphical Models*.

Shachter, R. D.; and Heckerman, D. 2010. Pearl Causality and the Value of Control. In R. Dechter, H. G.; and Halpern, J. Y., eds., *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. College Publications.

Tian, J.; and Pearl, J. 2001. Causal discovery from changes. In *Uncertainty in Artificial Intelligence (UAI)*.

Verma, T.; and Pearl, J. 1988. Causal Networks: Semantics and Expressiveness. In *Uncertainty in Artificial Intelligence (UAI)*.

Zhang, J.; Kumor, D.; and Bareinboim, E. 2020. Causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems* 33.

Zhang, L.; Wu, Y.; and Wu, X. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI International Joint Conference on Artificial Intelligence*.