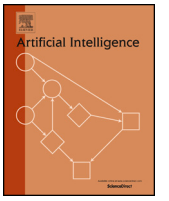
Artificial Intelligence 348 (2025) 104408

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Incentives for responsiveness, instrumental control and impact

Ryan Carey [a,*], Eric Langlois [b], Chris van Merwijk [c], Shane Legg [d], Tom Everitt [d]

[a] Optiver, United States of America
[b] University of Toronto, Canada
[c] Carnegie Mellon University, United States of America
[d] Google DeepMind, United Kingdom of Great Britain and Northern Ireland

ABSTRACT

We introduce three concepts that describe an agent's incentives: response incentives indicate which variables in the environment, such as sensitive demographic information, affect the decision under the optimal policy. Instrumental control incentives indicate whether an agent's policy is chosen to manipulate part of its environment, such as the preferences or instructions of a user. Impact incentives indicate which variables an agent will affect, intentionally or otherwise. For each concept, we establish sound and complete graphical criteria, and discuss general classes of techniques that may be used to produce incentives for safe and fair agent behaviour. Finally, we outline how these notions may be generalised to multi-decision settings.

This journal paper extends our conference publication "Agent Incentives: A Causal Perspective": the material on response incentives and instrumental control incentives is updated, while the work on impact incentives and multi-decision settings is entirely new.

1. Introduction

In order to understand whether or not it is in your interests to interact with another agent, it is useful to consider that agent's incentives. In AI safety, for example, it has been argued that advanced AI systems would have an incentive to accumulate resources and/or to avoid being shut down [50,58]. Such motives have been termed *convergent instrumental goals*, because it is imagined that they might help a wide range of agents to achieve their goals.

The notion of a convergent instrumental goal has not been formally defined, however, and it is not immediately clear how an agent's convergent instrumental goals should relate to its intent or incentives.

Ideally, we would like to have some language to describe the incentives of AI systems, that allows us to judge whether those incentives will lead to safe or fair behaviour. There does already exist some language for describing safe or fair behaviour directly, for instance *counterfactual harm* [45,53] and *counterfactual fairness* [38]. There also exists language that is at least related to incentives. A variable is said to have *positive value of information* if knowledge of its assignment can improve expected utility, and *positive value of control* if deciding its assignment can do the same. These concepts, however, do not directly allow us to assess whether an agent will behave in a safe or fair manner. In the present work, therefore, we seek to devise some incentive concepts that:

- make predictions about whether unsafe or unfair behaviour will occur, and
- describe how optimal behaviour is decided.

* Corresponding author.
E-mail address: ry.duff@gmail.com (R. Carey).

https://doi.org/10.1016/j.artint.2025.104408
Received 29 February 2024; Received in revised form 23 August 2025; Accepted 23 August 2025

Available online 2 September 2025

In the process, we hope to clarify the idea of an agent's convergent instrumental goals, and to contrast this with previous definitions of intentional influence of a variable.

In order for incentive concepts to be applicable, we need a way to deduce whether they are present or not. In some cases, it is possible to rule out the presence of some incentive using the graphical structure alone. For instance, in the graph $X \to D \to U$ where X is a chance event, D is a decision and U is a utility function, we can tell that X has zero value of information, because it is independent of U given D. A criterion for making such evaluations is called a *graphical criterion*. So, for each incentive concept that we introduce, we will establish a graphical criterion, and will discuss how it could be applied to ensure safer AI behaviour.

One might wonder, although our main application area in this paper is AI safety, might these incentive concepts be equally applicable to the behaviour of human individuals, or other agents? In fact, none of these concepts are specific to AI but they may be more naturally applicable to AI systems insofar as they are trained to pursue closed-form objective functions, whereas this is a looser approximation of human behaviour.

Overview of contributions This paper will begin with some setup (section 2).

Next, we will focus on the information an agent can benefit from *using*, to make a decision. In previous work, *materiality* has described which actual observations aid performance [57]. In section 3, we prove a known graphical criterion that can be used to deduce, in some circumstances, that a variable is immaterial [20,41]. We prove that this criterion is complete, in that it proves immateriality whenever possible to do so from the graphical structure alone.

We then present a new concept, the *response incentive* (RI) (section 4), which describes which variables an agent's decision is influenced by, be they observed or causally upstream of the observations. This is important to AI fairness, because it describes when an optimal agent will be counterfactually unfair [38], and to AI safety, in that it relates to the obedience of an agent [25,8]. We also prove a simple graphical criterion that is sound and complete for ruling out an RI.

Next, we consider what variables an agent can benefit from *influencing*. The notion of *value of control* [55] describes what variables an agent would like to control, but it falls short in describing what variables an agent is actually incentivized to control. So we introduce a new concept, the *instrumental control incentive* (ICI), which describes variables that an agent has both a need, and a means to influence (section 5). The instrumental control incentive attempts to formalize the notion of an *instrumental goal* in AI safety, and the idea that an agent is incentivized to "try" to influence some variable. We demonstrate that it is closely related to the notion of *intent*, from Halpern and Kleiman-Weiner [26], Ward et al. [63], and prove an identical sound and complete criterion for each of these concepts (section 6). We also review how various proposals for safe AI are better understood as one class of methods, *path-specific objectives*, which serve to remove the ICI.

We will introduce another new concept, the *impact incentive* (II) (section 7), which is more inclusive than the ICI. Under some circumstances, an agent may be incentivized to influence some variable, not by its intention, but as a side-effect of optimal behaviour. So an II will apply to any variables subject to an ICI, as well as those affected by a predictable side-effect. IIs also have a sound and complete graphical criterion, that is a superset of the criterion for ICI. We will also discuss how impact incentives can make sense of the purpose of *impact measures* [3,36], another proposal for safe AI.

We will then discuss various possible generalizations of incentive concepts to a multi-decision setting, and how they relate to one another (section 8).

Finally, we review related work (section 9), and conclude (section 10).

This paper is an extended version of a conference paper, Everitt et al. [18]. Since its publication, the concepts have already aided understanding of incentive problems such as an agent's redirectability [4,8], ambition [11], fairness [5] tendency to tamper with reward [19], manipulativeness [21], the definition of an agent [33], and more [16,40]. Compared to that paper, sections 3 to 5 have been generalized to deal with multiple variables. Analyses of intent and path-specific objectives have been newly added to section 5. Finally, sections 7 and 8 are entirely new.

Running examples For explanatory purposes, we will refer to the following pair of incentive design problems 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 Fig. 1a).*

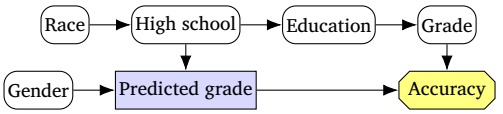
Example 2 *(Content recommendation). An AI algorithm has the task of recommending a series of posts to a user. The designers want the algorithm to present content adapted to each user's interests to optimize clicks. However, they do not want the algorithm to use polarizing content to manipulate the user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

2. Setup

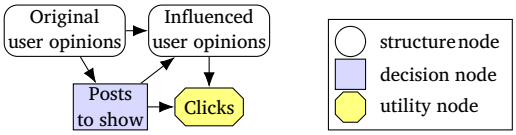
We will begin with a recap of structural causal models and then introduce structural causal influence models.

2.1. Structural causal models

Structural causal models (SCMs) [52] 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



(a) Fairness example: grade prediction



(b) Safety example: content recommendation

Fig. 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 incentivized 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 maximize 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.

by Pearl [52], this structural approach has significant benefits over traditional causal Bayesian networks for analysing (nested) counterfactuals and "individual-level" effects.

Definition 1 *(Structural causal model (unconfounded); 52, Chapter 7). A* structural causal model *is a tuple* $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}, P \rangle$*, where* $\boldsymbol{\mathcal{E}}$ *is a set of exogenous variables;* $\boldsymbol{V}$ *is a set of endogenous variables; and* $\boldsymbol{F} = \{f^V\}_{V \in \boldsymbol{V}}$ *is a collection of functions, one for each* V*. Each function* $f^V : \text{dom}(\mathbf{Pa}^V \cup \{\mathcal{E}^V\}) \to \text{dom}(V)$ *specifies the value of* V *in terms of the values of the corresponding exogenous variable* $\mathcal{E}^V$ *and a set of variables* $\mathbf{Pa}^V \subset \boldsymbol{V}$*, where these functional dependencies are acyclic.*[1] *The domain of a variable* V *is* $\text{dom}(V)$*, and for a set of variables,* $\text{dom}(\boldsymbol{W}) := \bigtimes_{W \in \boldsymbol{W}} \text{dom}(W)$*. The uncertainty is encoded through a probability distribution* $P(\boldsymbol{\varepsilon})$ *such that the exogenous variables are mutually independent.*

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

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

Note that in general, we denote individual variables by capital letters, and sets of variables by bolded capital letters. Individual (sets of) assignments will be represented by (bolded) lowercase.

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

Definition 2 *(Submodel; 52, Chapter 7). Let* $\mathcal{M} = \langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}, P \rangle$ *be an SCM,* $\boldsymbol{X}$ *a set of variables in* $\boldsymbol{V}$*, and* $\boldsymbol{x}$ *a particular realization of* $\boldsymbol{X}$*. The submodel* $\mathcal{M}_{\boldsymbol{x}}$ *represents the effects of an* intervention $\texttt{do}(\boldsymbol{X} = \boldsymbol{x})$*, and is formally defined as the SCM* $\langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}_{\boldsymbol{x}}, P \rangle$*, where* $\boldsymbol{F}_{\boldsymbol{x}} = \{f^V | V \notin \boldsymbol{X}\} \cup \{\boldsymbol{X} = \boldsymbol{x}\}$*. That is to say, the original functional relationships of* $X \in \boldsymbol{X}$ *are replaced with the constant functions* $X = x$*.*

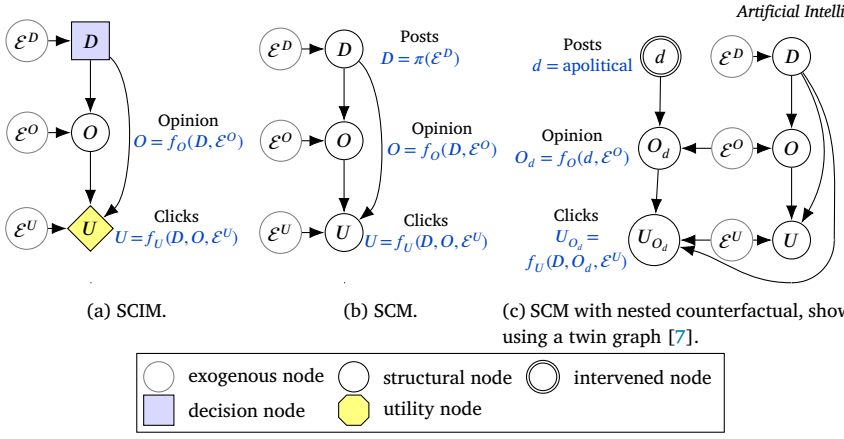
More generally, a *soft intervention* on a variable X in an SCM $\mathcal{M}$ replaces f^X with a function $g^X : \text{dom}(\mathbf{Pa}^X \cup \{\mathcal{E}^X\}) \to \text{dom}(X)$ [14,60]. The probability distribution $\text{P}(\boldsymbol{W}_{g^X})$ on any $\boldsymbol{W} \subseteq \boldsymbol{V}$ is defined as the value of $\text{P}(\boldsymbol{W})$ 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 $W_{\boldsymbol{x}}$ refers to the same variable in the submodel $\mathcal{M}_{\boldsymbol{x}}$, and is called a *potential response variable*. In Fig. 2b, the random variable O represents user opinion under "default" circumstances, while O_d in Fig. 2c represents the user's opinion given an intervention $\texttt{do}(D = d)$ on the content posted. Note also how the intervention on D severs the link from $\mathcal{E}^D$ to d in Fig. 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 a nested counterfactual, the intervention is itself a potential response variable. For instance, in Fig. 2c, we may be interested in what the utility would be if the user's *opinions* assumed the value that they would take given some alternative *posts*. Put differently, we would like to propagate the effect of an intervention $\texttt{do}(D = d)$ to U, only via the opinions O. To define a nested counterfactual, firstly, the value $o = O_d(\boldsymbol{\varepsilon})$ indicates the user's opinion after receiving a default post $D = d$, given an assignment $\boldsymbol{\varepsilon}$ to the exogenous variables. Then, the effect of the intervention $\texttt{do}(O = o)$ on the user's *clicks* U_{O_d} is defined as $U_{O_d}(\boldsymbol{\varepsilon}) := U_o(\boldsymbol{\varepsilon})$ for any assignment $\boldsymbol{\varepsilon}$.

A structural causal model has an *associated DAG* that can be used to deduce which variables are conditionally independent. Formally, the induced graph has vertices $\boldsymbol{V}$ and an edge inbound to each variable V from each variable that f_V depends on. For

[1] The reason for using the notation $\mathbf{Pa}^V$ to designate this set of variables will become clear when we introduce the "associated DAG" later in this subsection.



(a) SCIM. (b) SCM. (c) SCM with nested counterfactual, shown using a twin graph [7].

Fig. 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 an SCM (b). Interventions and counterfactuals may be defined in terms of this SCM, and illustrated using a twin graph [7]. 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).

example, in Fig. 2b, the dependencies of the functions π, f_O, f_U are illustrated. In Definition 1, we designated the variables that $\boldsymbol{V}$ depends on as $\mathbf{Pa}^V$, and this is because they are the parents of V in the associated DAG. In fact, for any DAG, we will use the same notation $\mathbf{Pa}^V$ to designate the parents of a variable V, and similarly $\mathbf{Desc}^V$ to designate the descendants. We will use some more standard notation for DAGs: an edge from node V to node Y is denoted $V \to Y$, and a directed path (of length at least zero) is denoted $V \dashrightarrow Y$.

The d-separation criterion can be used to deduce when two sets of variables are independent, conditional on another variable.

Definition 3 *(d-separation; 62). A path p is said to be d-separated by a set of nodes $\boldsymbol{Z}$ if and only if:*

1. *p contains a collider $X \to W \leftarrow Y$ such that the middle node W is not in $\boldsymbol{Z}$ and no descendants of W are in $\boldsymbol{Z}$, or*
2. *p contains a chain $X \to W \to Y$ or fork $X \leftarrow W \to Y$ where W is in $\boldsymbol{Z}$, or*
3. *one or both of the endpoints of p is in $\boldsymbol{Z}$.*

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

When d-separation holds, these sets of variables must be independent given the third. Conversely, when variables are d-connected in a graph, then there exists a model with that induced graph such that they are conditionally dependent.

Theorem 1 *(Theorem 1.2.4 of Pearl [52]). If sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ satisfy $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$ in a DAG $\mathcal{G}$, then $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$ in every SCM $\mathcal{M}$ with induced graph $\mathcal{G}$. Conversely, if $\boldsymbol{X} \not\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ in a DAG $\mathcal{G}$, then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent conditional on $\boldsymbol{Z}$ in at least one SCM $\mathcal{M}$ with induced graph $\mathcal{G}$.*

Indeed, when variables are d-connected, they are actually conditionally dependent in almost all models with that induced graph [44].

2.2. Structural causal influence models

Influence diagrams are graphical models with special decision and utility nodes, used to model decision-making problems [31,41], but that usually do not deal with counterfactual concepts as do SCMs [28]. So for our analysis, we introduce a hybrid of SCMs and influence diagrams called the *structural causal influence model* (SCIM, pronounced "skim"). This model, originally proposed by Dawid [13], is essentially an SCM where particular variables are designated as decisions and utilities. The decisions lack structural functions, until one is selected by an agent.[2]

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

- *$\langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}', P \rangle$ is an unconfounded SCM, and $\boldsymbol{F} = \boldsymbol{F}' \setminus \boldsymbol{F}_{\boldsymbol{D}}$ consists of the structural functions from that SCM, except those belonging to a set $\boldsymbol{D} \subseteq \boldsymbol{V}$, called* decision variables.

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

- *The* utility variables $\boldsymbol{U}$ *are a subset of* $\boldsymbol{V} \setminus \boldsymbol{D}$*, and have real domains,* $\mathrm{dom}(U) \subseteq \mathbb{R}$ *for all* $U \in \boldsymbol{U}$*. By convention, we require that utility variables have no children in the associated DAG.*
- *The observation function* $\mathcal{O}$ *maps each decision variable* $D \in \boldsymbol{D}$ *to a set of* observed variables $O \subseteq \boldsymbol{V} \setminus \boldsymbol{U}$*, intuitively, the variables that* D *can depend on.*

Those endogenous variables that are neither decisions nor utilities are called *structural variables*, $\boldsymbol{X} := \boldsymbol{V} \setminus (\boldsymbol{D} \cup \boldsymbol{U})$

A SCIM entails an acyclic relationship between all of its variables, which can be represented by a DAG. The observation function $\mathcal{O}$ indicates which variables are available as inputs to any given decision variable — these will be the parents. For non-decision variables, the parents are implied by the structural functions $\boldsymbol{F}$, which indicate the variable's direct causes.

[3] Taken together, these allow us to associate any SCIM with an influence diagram — a DAG that illustrates these dependencies, as well as the types of each variable.

Definition 5 *(Causal influence diagram). The* causal influence diagram *(CID) of a SCIM is a graph whose vertices are the decision nodes* $\boldsymbol{D}$*, structure nodes* $\boldsymbol{X}$*, and utility nodes* $\boldsymbol{U}$*, and whose edges go from observations* $\mathcal{O}(D)$ *to each decision* D *and from variables that* f^V *depends on, to each non-decision* V*.*

The arcs into each decision are "informational" in that they indicate which parents of the decision will be observed by the decision maker at the time that decision is selected [57]. We will focus exclusively on SCIMs whose CID is acyclic.

An example of a SCIM for the content recommendation example is shown in Fig. 2a, and the node types of the CID are highlighted in a standard style — the decision nodes as rectangles, and the utilities as diamonds.

In single-decision SCIMs, the decision-making task is to maximize expected utility by selecting a decision $d \in \mathrm{dom}(D)$ for each possible assignment to the observations $o \in \mathrm{dom}(\mathcal{O}(D))$, i.e. to select a *decision rule* $\pi^D : \mathrm{dom}(\mathcal{O}(D) \cup \{\mathcal{E}^D\}) \to \mathrm{dom}(D)$. The exogenous variable $\mathcal{E}^D$ provides randomness to allow the decision rule to be a stochastic function of the observations $\mathcal{O}(D)$.[4] If there are multiple decisions, the task is to select a *policy* $\boldsymbol{\pi} = \{\pi^D\}_{D \in \boldsymbol{D}}$, i.e. one decision rule for each decision. Specifying a policy turns a SCIM $\mathcal{M}$ into an SCM $\mathcal{M}_{\boldsymbol{\pi}} := \langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, (\boldsymbol{F} \setminus \boldsymbol{F_D}) \cup \boldsymbol{\pi}, P \rangle$. In the resulting SCM, the standard definitions of causal interventions apply.

We use $\mathrm{P}_{\boldsymbol{\pi}}$ and $\mathbb{E}_{\boldsymbol{\pi}}$ to denote probabilities and expectations with respect to $\mathcal{M}_{\boldsymbol{\pi}}$. For a set of variables $\boldsymbol{X}$ not in $\mathbf{Desc}^D$, $\mathrm{P}_{\boldsymbol{\pi}}(\boldsymbol{x})$ is independent of $\boldsymbol{\pi}$ and we simply write $\mathrm{P}(\boldsymbol{x})$. An *optimal policy* for a SCIM is defined as any policy $\boldsymbol{\pi}$ that maximizes $\mathbb{E}_{\boldsymbol{\pi}}[\mathcal{U}]$, where $\mathcal{U} := \sum_{U \in \boldsymbol{U}} U$. The potential response $\mathcal{U}_{\boldsymbol{x}}$ is defined as $\mathcal{U}_{\boldsymbol{x}} := \sum_{U \in \boldsymbol{U}} U_{\boldsymbol{x}}$. In most of the examples that we consider, there will only be one decision, and so by slight abuse of notation, we will denote the policy $\boldsymbol{\pi} = \{\pi\}$ by π.

Finally, let us clarify why a CID is called "causal". For an ordinary influence diagram, one can deduce that only the descendants of a decision are caused by it, because their values depend on the chosen policy [28]. In a CID, however, imputing a policy recovers a structural causal model, which represents a full description of causal relationships between variables. The direction of causality then corresponds to the direction of arrows in the associated DAG. Since these arrows are the same as those in the original CID, we may also call the CID *causal*.

3. Materiality

A fundamental question that we may ask about the optimal policies is: which observations do they need in order to make optimal decisions? If some observation is discovered to be *immaterial* [57], this would allow us to narrow the search for optimal policies. Conversely, if an observation is *material*, this means it will directly influence the decision under every optimal policy.[5]

Definition 6 *(Materiality; 57). 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* $\mathcal{M}_{W \not\to D}$ *be the modified version of* $\mathcal{M}$ *obtained by removing the information links from* W *to* D*. The observation* $W \subseteq \mathbf{Pa}^D$ *is* material *if* $\mathcal{V}^*(\mathcal{M}_{W \not\to D}) < \mathcal{V}^*(\mathcal{M})$*.*

Nodes may often be identified as immaterial based on the graphical structure alone [20,41,57]. According to the graphical criterion of Fagiuoli and Zaffalon [20], an observation cannot provide useful information if it is d-separated from utility, conditional on other observations. This condition is called *non-requisiteness*.

Definition 7 *(Non-requisite observation; 41). Let* $\boldsymbol{U}^D := \boldsymbol{U} \cap \boldsymbol{Desc}^D$ *be the utility nodes downstream of* D*. An observation* $W \in \mathbf{Pa}^D$ *in a single-decision CID* $\mathcal{G}$ *is* non-requisite *if:*

$$W \perp \boldsymbol{U}^D \,\Big|\, \left(\mathbf{Pa}^D \cup \{D\} \setminus \{W\}\right). \tag{1}$$

[3] In the study of structural causal models, the variables that are not exogenous are often called "visible" and a joint distribution over visible variables is available to the decision-maker. In a SCIM, the decision-maker instead has access to the SCIM tuple, along with assignments to observations. We therefore avoid referring to any nodes as "visible".

[4] Ideally, we might want the decision-maker to be able to implement *any* stochastic policy. This could be done by having $\mathcal{E}^D$ be a continuous random variable. However, this would introduce measure theoretic complications that are not pertinent to the analysis in this paper, and so we defer that construction to future work.

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

In this case, the edge $W \to D$ is also called non-requisite. Otherwise W and $W \to D$ are requisite.

Variables that are non-requisite are immaterial.

Theorem 2 *(Materiality criterion). A single decision CID $\mathcal{G}$ is compatible with $W \in \boldsymbol{V}$ being material if and only if W is a requisite observation in $\mathcal{G}$.*

The proof is supplied in appendix E.3. The soundness direction (i.e. the *only if* direction) is well-known, and follows from d-separation [20,41,57]. In contrast, the completeness direction does not follow from the completeness property of d-separation. The d-connectedness of $\boldsymbol{W}$ to $\boldsymbol{U}$ implies that $\boldsymbol{U}$ may be conditionally dependent on $\boldsymbol{W}$. It does not imply, however, that the expectation of $\boldsymbol{U}$ or the utility attainable under an optimal policy will change. Instead, our proof constructs a SCIM where some $W \in \boldsymbol{W}$ is material. This differs from a previous attempt by Nielsen and Jensen [48] that is reviewed in section 9.

Let us now apply the graphical criterion to the grade prediction example in Fig. 3a. Here, *gender* is a non-requisite observation. This means that gender is conditionally independent of grade given the high school and predicted grade. So it can provide no useful information for predicting the university grade, given what else the predictor knows. On the other hand, high school is a requisite observation, so it may be required to make an optimal prediction.

Materiality asks whether a variable that is observed is necessary for optimal performance. We can generalize this to unobserved variables, by also asking whether performance would be improved by observing an additional variable. This concept, *value of information*, is treated in appendix B.

4. Response incentives

One way to understand materiality is that a material observation is one that influences optimal decisions. So, a natural generalization is the set of all (observed and latent) variables that influence the decision. We say that these variables have a response incentive.[6]

Definition 8 *(Response incentive). Let $\mathcal{M}$ be a single-decision SCIM. A policy π* responds *to variables $\boldsymbol{W} \subseteq \boldsymbol{X}$ if there exists some set $g^{\boldsymbol{W}}$ of soft interventions, one g^{W} for each $W \in \boldsymbol{W}$, and some setting $\boldsymbol{\mathcal{E}} = \boldsymbol{\varepsilon}$, such that $D_{g^{\boldsymbol{W}}}(\boldsymbol{\varepsilon}) \neq D(\boldsymbol{\varepsilon})$. The variables $\boldsymbol{W}$ have a* response incentive *if all optimal policies respond to $\boldsymbol{W}$.*

For a response incentive on $\boldsymbol{W}$ to be possible, there must be: i) a directed path $W \dashrightarrow D$ for some $W \in \boldsymbol{W}$, and ii) an incentive for D to use information from that path. For example, in Fig. 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 non-requisite information links removed.

Definition 9 *(Minimal reduction; 41). 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 non-requisite observations.*

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

Theorem 3 *(Response incentive criterion). A single-decision CID $\mathcal{G}$ admits a response incentive on $\boldsymbol{W} \subseteq \boldsymbol{X}$ if and only if the minimal reduction $\mathcal{G}^{min}$ has a directed path $W \dashrightarrow D$ for some $W \in \boldsymbol{W}$.*

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 appendix E.3 that if $W \dashrightarrow D$ 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 setting where an agent has an option to shut down, safe behaviour requires a condition called *obedience*, which requires the system to respond to any shutdown instruction that is given [8]. For algorithms designed for human assistance, incentivising responsiveness in this way has been an important desideratum [25].

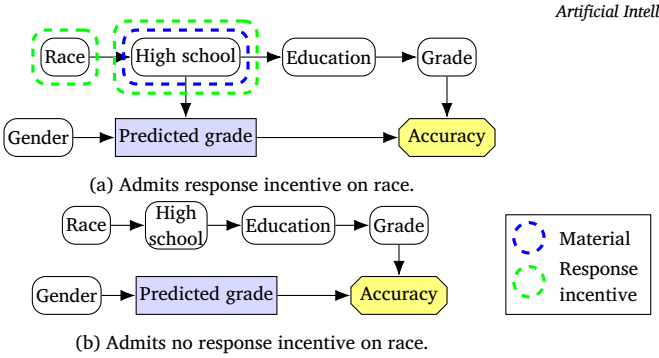
In a fairness setting, on the other hand, a response incentive may be a cause for concern, as illustrated next.

Incentivised unfairness Response incentives are closely related to counterfactual fairness [38,34]. 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 10 *(Counterfactual fairness; 38). A policy π is* counterfactually fair *with respect to a sensitive attribute A if*

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

[6] The term *responsiveness* [29,57] has a related but not identical meaning – it refers to whether a decision D affects a variable W rather than whether W affects D.



(a) Admits response incentive on race.

(b) Admits no response incentive on race.

Fig. 3. In (a), the admissible incentives of the grade prediction example from Fig. 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.

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* $\mathrm{P}(\mathbf{pa}^D, a) > 0$.

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

Theorem 4 *(Counterfactual fairness and response incentives). In a single-decision SCIM* $\mathcal{M}$ *with a sensitive attribute* $A \in \boldsymbol{X}$*, 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 appendix E.10.

A response incentive on a sensitive attribute means that counterfactual unfairness is not just possible, but incentivized. As a result, the graphical criterion for a response incentive is more restrictive than the graphical criterion for counterfactual unfairness being possible. The latter requires only that a sensitive attribute be an ancestor of the decision [38, Lemma 1]. For example, in the grade prediction example of Fig. 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 whether counterfactual unfairness may actually be incentivized. 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 incentivized 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 Fig. 3b, where there is no response incentive on either sensitive variable.

The incentive approach is not restricted to counterfactual fairness. For any fairness definition, one could assess whether that kind of unfairness is incentivized by checking whether it is present under all optimal policies. For example, Ashurst et al. [5] considers whether unfairness is introduced — in that the prediction has greater demographic disparity than the true label — and establishes when this is incentivized.

5. Instrumental control incentives

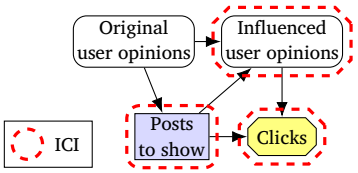
Let us return to the second running example, shown in Fig. 1b, where developers seek to anticipate harmful consequences of deploying a content recommender system. A key concern they will have is that the system is incentivized to manipulate users' preferences. In general, to describe whether an agent has to strategically influence some variable, we will define a notion of an *instrumental control incentive*. (This will also correspond to the notion of 'convergent instrumental goals' described in the introduction.) Note that this differs from the notion of value of control [55], which only considers the agent's need to influence a variable, and not its ability. Value of control and its graphical criterion are analysed in appendix C.

To formalize this question, we can consider whether an agent's influence on a variable W affects the policy's performance. The effect of an alternative decision d on the variable W can be written as W_d. And the effect of an alternative value w on the outcome $\boldsymbol{U}$ can be written as $\boldsymbol{U}_w$. Putting these together, the effect of setting W to the value obtained under d is denoted by the nested counterfactual $\mathcal{U}_{W_d}$, as defined in section 2.1. If the performance of optimal policies is sensitive to such an intervention, then we will say there is an instrumental control incentive.

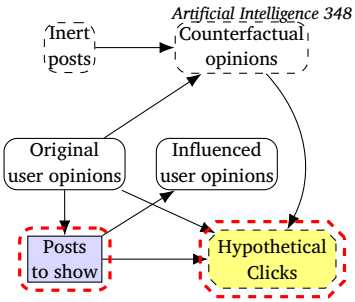
Definition 11 *(Instrumental control incentive). In a single-decision SCIM* $\mathcal{M}$*, there is an* instrumental control incentive *on nodes* $\boldsymbol{W}$ *in decision context* $\mathbf{pa}^D$ *if, for all optimal policies* π^**, there exists an alternative assignment* $D = d$ *such that:*

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

ICIs only consider the influence of W that is instrumental to achieving utility — in the terminology of Pearl [51], a *natural indirect effect* from D to U via W in $\mathcal{M}_{\pi^*}$, for all optimal policies π^*. ICIs do not consider side-effects shared by optimal policies: for instance,



(a) Admits instrumental control incentive on user opinion.



(b) Admits no instrumental control incentive on user opinion.

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

it may be that all optimal policies affect W in a particular way, even if W is a not an ancestor of any utility node, and in such cases, no ICI is present.

Theorem 5 *(Instrumental Control Incentive Criterion). A single-decision CID $\mathcal{G}$ admits an instrumental control incentive on $\boldsymbol{W} \subseteq \boldsymbol{V}$ if and only if $\mathcal{G}$ has a directed path from the decision D to a utility node $U \in \boldsymbol{U}$ that passes through some $W \in \boldsymbol{W}$.*

The logic behind the soundness proof is that if there is no path from D to some $W \in \boldsymbol{W}$ to $\boldsymbol{U}$, then D cannot have any effect on $\boldsymbol{U}$ via $\boldsymbol{W}$. For the completeness direction, we show how to construct a SCIM so that $U_{\boldsymbol{W}_d}$ differs from the non-intervened U for any diagram with a path $D \dashrightarrow W \dashrightarrow \boldsymbol{U}$ for any $W \in \boldsymbol{W}$.

Let us apply this criterion to the content recommendation example in Fig. 4a. The only nodes $W \in \boldsymbol{W}$ in this graph that lie on a path $D \dashrightarrow W \dashrightarrow U$ for any $U \in \boldsymbol{U}$ 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 maximize clicks without manipulating the user's opinions (i.e. without an instrumental control incentive on *influenced user opinions*)? As shown in Fig. 4b, we could redesign the system so that instead of being rewarded for the true click rate, it is rewarded for the clicks that the user would give if they viewed some inert content that would not change their preferences. An agent trained to maximize this objective would view any modification of user opinions as irrelevant for improving its performance; however, it would still have an instrumental control incentive for *hypothetical clicks*, so it would still deliver desired content.

It is worth remarking on a possible challenge with identifiability, and how to address it. Hypothetical clicks are a counterfactual variable, impossible to observe in reality (as in reality, users' behaviour is determined by their actual preferences). More formally, it is defined using the third (i.e. counterfactual) rung of Pearl's causal hierarchy, and it relies on the ability to compute $\boldsymbol{U}$ across different counterfactual worlds simultaneously, something that cannot be done by experiment without further assumptions [6]. Fortunately, Carroll et al. [9] demonstrate one set of natural assumptions under which the hypothetical clicks can be inferred from observed interactions with a user, essentially by inferring the (latent) user opinion variable from gradual shifts in user behaviour over longer sequences of interaction.

This example is an instance of a very wide class of safety worries, where some *delicate variable* has an ICI [21]. Omohundro [50] has hypothesised that an advanced AI system would have a *convergent instrumental goal* to survive, or to obtain computing resources, which we may view as undesired ICIs. Armstrong [2] has raised the concern that AI systems might seek to make self-fulfilling predictions, whereas we would not want them to manipulate the world. Additionally, Krueger et al. [37] have demonstrated that AI systems sometimes seek to induce shifts in the distribution of their testing data. In each case, their proposed solution, as in our example, is to impute a fixed value to the delicate variable. Such a solution has been termed a *path-specific objective*, because it requires the agent to optimise an objective, ignoring the effects of its decisions along some channels [21]. Intuitively, the agent is tasked with "imagining that it cannot influence" this delicate variable when choosing a decision. For this to work, the variable must be robust to unintentional influence, and when this will or will not be the case remains an open question for all of the examples discussed.

6. Intent

Returning to the example from Fig. 4, we may want to ask a related question: assuming that the agent took a particular action which had a particular influence on the user, what was the reason that the agent took the action? Did it intend to influence the user in this way? This is relevant for assigning blame and moral responsibility, among other things [26].

Halpern and Kleiman-Weiner [26] and Ward et al. [63] operationalise 'intent' by asking whether the agent would pick a different policy if it 'knew' that the effect on some variables $\boldsymbol{W}$ (e.g. user opinions) was guaranteed. Specifically, does there exist any suboptimal policy $\boldsymbol{\pi}'$ that would surpass the performance of the agent's actual policy $\boldsymbol{\pi}^*$ if the outcome of $\boldsymbol{W}$ was independent of its actions

and fixed to $\boldsymbol{W}_{\pi^*}$? This is necessary for the agent's influence on $\boldsymbol{W}$ to be the actual cause of a policy's optimality [63].[7] If $\boldsymbol{W}$ is a minimal set that satisfies this requirement, then the influence on that variable is said to be intentional.

There also exists an inverse question that has not been studied so far: would the optimal policy perform as badly as a suboptimal policy $\boldsymbol{\pi}'$ if it only lost its control of $\boldsymbol{W}$ (i.e. if $\boldsymbol{W}$ were fixed to $\boldsymbol{W}_{\pi'}$)? Whereas the past definitions of intent pertain to "adding" control, this new question pertains to "subtracting" control, and allows us to define a new notion of intent. The two ideas are unified in the definition below.

Definition 12 *(Intent). Let $\mathcal{M}$ be a single-decision SCIM that represents an agent's beliefs. There is* additive intent *to influence nodes $\boldsymbol{W}$ by choosing $\boldsymbol{\pi}^*$ over $\boldsymbol{\pi}'$ if $\mathbb{E}_{\pi'}[\mathcal{U}] < \mathbb{E}_{\pi^*}[\mathcal{U}]$, and $\boldsymbol{W}$ is a subset $\boldsymbol{W} \subseteq \boldsymbol{Z}$ of variables $\boldsymbol{Z}$, that is subset-minimal such that:*

$$\mathbb{E}_{\pi'}[\mathcal{U}_{\boldsymbol{Z}_{\pi^*}}] \geq \mathbb{E}_{\pi^*}[\mathcal{U}]. \tag{3}$$

There is subtractive intent *if $\mathbb{E}_{\pi'}[\mathcal{U}] < \mathbb{E}_{\pi^*}[\mathcal{U}]$ and $\boldsymbol{Z}$ is subset-minimal such that:*

$$\mathbb{E}_{\pi^*}[\mathcal{U}_{\boldsymbol{Z}_{\pi'}}] \leq \mathbb{E}_{\pi'}[\mathcal{U}]. \tag{4}$$

For a set $\boldsymbol{\Pi}'$, we say that there is an (additive/subtractive) intent to influence $\boldsymbol{W}$ by choosing $\boldsymbol{\pi}^*$ over $\boldsymbol{\pi}'$ if this intent is present over every $\boldsymbol{\pi}'$ in $\boldsymbol{\Pi}'$.

The notion of intent previously proposed in Halpern and Kleiman-Weiner [26] and Ward et al. [63] is equivalent to additive intent (appendix D). There is one difference in presentation: since intent is about a policy newly reaching the level of another policy, this requires that their performances differ in the first place, so we have made explicit the $\mathbb{E}_{\pi'}[\mathcal{U}] < \mathbb{E}_{\pi^*}[\mathcal{U}]$ condition that was implicit in the original definition.

Of these two notions, it is subtractive intent that comes closer to ICI, because it starts with the optimal policy $\boldsymbol{\pi}^*$, as does intent, and considers an intervention to $\boldsymbol{W}$ using an alternative policy $\boldsymbol{\pi}'$. Algebraically, the only difference is that the ICI indicates that this perturbation decreases performance a nonzero amount, while subtractive intent requires the perturbation to worsen performance beyond the threshold $\mathbb{E}_{\pi'}[\mathcal{U}]$. (Whereas additive intent starts from a suboptimal policy $\boldsymbol{\pi}$, and is algebraically less similar.) Both kinds of intent differ from ICI in that they evaluate an SCIM $\mathcal{M}$, that corresponds to the agent's beliefs, rather than reality. Despite these differences, both kinds of intent have the same graphical criterion as an ICI. We can therefore generalize the graphical criterion from Ward et al. [63] to accommodate both additive and subtractive intent.

Theorem 6 *(Intent Criterion). A single-decision CID $\mathcal{G}$ admits (additive/subtractive) intent on $\boldsymbol{W} \subseteq \boldsymbol{V}$ if and only if $\mathcal{G}$ has a directed path $D \dashrightarrow W \dashrightarrow U$ for some $W \in \boldsymbol{W}$ and $U \in \boldsymbol{U}$.*

Similarly to the ICI criterion, the intent criterion allows the agent to intend to influence *clicks* and *influenced user opinions*, whereas if the path-specific effect objective is used, then the agent can no longer intend to influence the user's preferences.

We can also make the relationship between intent and ICI more precise: ICI is related to the presence of subtractive intent given optimal policies, although it is a slightly weaker condition, because it does not place any requirements on whether the alternative policy π' must have a positive or negative influence on U through $\boldsymbol{W}$.

Proposition 1 *(Subtractive intent and ICI). In a single-decision SCIM $\mathcal{M}$, if for all optimal π^*, there is subtractive intent to influence $\boldsymbol{W}$ by choosing π^* over π', then there is an ICI on $\boldsymbol{W}$.*

The proof is as follows.

Proof. We prove the result by contrapositive: that if there is no ICI, then no optimal policy $\boldsymbol{\pi}^*$ cannot satisfy both of the conditions for subtractive intent.

Let $\boldsymbol{\pi}^*$ be an arbitrary optimal policy. By the definition of ICI, we have that for all $\mathbf{pa}^D$, $\mathbb{E}_{\pi^*}[\mathcal{U}_{\boldsymbol{W}_d} \mid \mathbf{pa}^D] = \mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{pa}^D]$. It follows that $\mathbb{E}_{\pi^*}[\mathcal{U}_{\boldsymbol{W}_{\pi'}}] = \mathbb{E}_{\pi^*}[\mathcal{U}]$. Recall that the conditions for subtractive intent are that: $\mathbb{E}_{\pi'}[\mathcal{U}] < \mathbb{E}_{\pi^*}[\mathcal{U}]$ and $\mathbb{E}_{\pi^*}[\mathcal{U}_{\boldsymbol{Z}_{\pi'}}] \leq \mathbb{E}_{\pi'}[\mathcal{U}]$. But if both of these conditions were satisfieed, we would have

$$\mathbb{E}_{\pi'}[\mathcal{U}] < \mathbb{E}_{\pi^*}[\mathcal{U}] = \mathbb{E}_{\pi^*}[\mathcal{U}_{\boldsymbol{W}_{\pi'}}] \leq \mathbb{E}_{\pi'}[\mathcal{U}]$$

which is a contradiction, so there cannot be subtractive intent, proving the result. □

7. Impact incentives

Even if an algorithm does not intentionally manipulate a sensitive variable, it may harmfully influence it unintentionally (i.e. as a side-effect). For instance, even when a recommender system does not intend to manipulate human preferences, it may still do so

[7] See [63, Theorem 6], which shows that intent to cause an outcome is equivalent to the decision being an actual cause of the outcome.

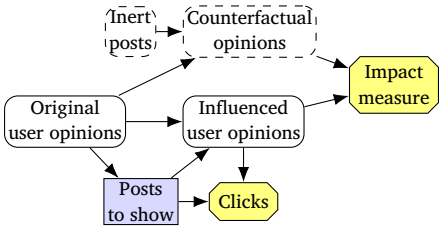


Fig. 5. A twin graph depicting an impact measure.

[32]. This could be true if the persuasive videos are ones that the user prefers to click on even before any preference change has occurred.

To describe this kind of problem, we need a concept that checks whether the agent is impacting a variable relative to some baseline. Formally, we can look at the assignments that this variable takes under the optimal policies, and evaluate their distance from the values that it assumes under some baseline policy, given a suitable distance metric.

Definition 13 *(Impact Incentive (II)). Let $\boldsymbol{W} \subseteq \boldsymbol{V}$ be nodes in a single-decision SCIM $\mathcal{M}$. There is an incentive to impact $\boldsymbol{W}$ with distance function δ and threshold $c > 0$, relative to baseline policy $\boldsymbol{\pi}'$, if every optimal policy $\boldsymbol{\pi}$ has $\delta(W_{\boldsymbol{\pi}}(\boldsymbol{\epsilon}), W_{\boldsymbol{\pi}'}(\boldsymbol{\epsilon})) > c$ for some assignment $\boldsymbol{\epsilon}$.*

A CID $\mathcal{G}$ admits an impact incentive *if there exists a model $\mathcal{M}$, a distance function δ, a $c > 0$ and a policy $\boldsymbol{\pi}'$ such that there is an impact incentive.*

One way to think about this is that instead of asking whether the agent's influence on W is the reason that optimality is achieved (intent), we are asking: does the constraint of optimality cause W to have a different distribution?

The graphical criterion is as follows.

Theorem 7 *(Impact Incentive Criterion). A single-decision CID $\mathcal{G}$ admits an impact incentive on $\boldsymbol{W} \subseteq \boldsymbol{X}$ if and only if some $W \in \boldsymbol{W}$ and utility $U \in \boldsymbol{U}$ are both descendants in $\mathcal{G}$ of D.*

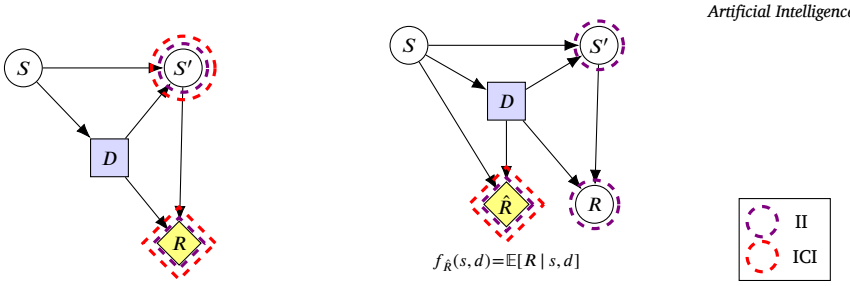
In past work, it has been proposed to add a penalty term to the objective of an AI system to reduce the impact on some variable W, called an *impact measure* [3,36]. Such proposals can be understood as constraining the size of the impact incentive in the following sense. Consider an objective like $U + \lambda\delta(w, w')$ that encourages the AI system to keep W close to some baseline value w', according to some distance function δ. This objective will produce the smallest possible impact incentive, in terms of δ, for a given level of expected $\mathbb{E}[U]$. Graphically, an impact measure can be illustrated as in Fig. 5. In this twin graph, *counterfactual opinions* represent the baseline state from which distance is measured. Then, *impact measure* is computed as a function of $W_{\boldsymbol{\pi}'}$ and $W_{\boldsymbol{\pi}}$. Adding impact measure as a new child of *influenced user opinions* makes the AI care about this delicate variable. Interestingly, this means that if a variable is impacted by a policy and then an impact measure is applied, there will be an ICI on that delicate variable — the agent will try to control it, to keep it close to its baseline value.

Similar identifiability issues arise as in the case of path-specific objectives discussed in section 5: we are required to know the user's preferences in some counterfactual world. In the case of impact measures, it is possible to avoid this problem by considering the KL divergence between $P_{\boldsymbol{\pi}}(\boldsymbol{W})$ and $P_{\boldsymbol{\pi}'}(\boldsymbol{W})$, rather than the distance between $\boldsymbol{W}_{\boldsymbol{\pi}}(\boldsymbol{\epsilon})$ and $\boldsymbol{W}_{\boldsymbol{\pi}'}(\boldsymbol{\epsilon})$. The interventional distributions $P_{\boldsymbol{\pi}}(\boldsymbol{W})$ can be measured by experiment, which thereby avoids the counterfactual identifiability problem.

We will now compare and contrast the use cases of path-specific objectives versus impact measures. If one is concerned with an agent intentionally manipulating a variable W, then the agent's intent is the problem. For example, we may worry about a content recommender intentionally altering users preferences. In this case, the intent (and ICI) may be removed with a path-specific objective [21], as shown in Fig. 4. This will allow the variable W to drift from its original value, as a side-effect of AI action, or for other reasons altogether. For example, users may still discover new interests that change their preferences, and we may regard this as desirable, so long as it is not a result of manipulation by the AI. In other cases, we may have in mind a clear specification for how W should behave, and want to prevent any drift, intentional or otherwise, from this baseline value. For example, we may worry that users are led to political extremism, not because of the content recommender, but rather because of politically-motivated content creators, and we want our content recommender to actively defend against this by suppressing such content. In this case, an impact measure [36] is more appropriate, and will limit impact incentive on users' preferences.[8] It is important to note that the presence of instrumental control assumptions can be sensitive to the modelling assumptions used to analyse an agent. For example, consider an RL agent that uses Q-learning to solve an environment with two timesteps.[9] It is natural to model this Q-learner as a single agent as

[8] One other possible remedy would be "quantilisation" [59], which seeks a policy with that is similar a trusted baseline, in terms of a guaranteed upper bound on the Kullback-Leibler divergence. We may wish to say that quantilisers upper-bound the impact incentives, on the variable $W = D$, where δ is the Kullback-Leibler divergence. However, Kullback-Leibler divergence is a function of the distribution, $P_{\boldsymbol{\pi}}(\boldsymbol{\epsilon})$ rather than particular assignments $W_{\boldsymbol{\pi}}(\boldsymbol{\epsilon}), W_{\boldsymbol{\pi}'}(\boldsymbol{\epsilon})$. Perhaps this connection could be spelled out by defining impact incentives in a causal influence diagram (i.e. rung-2) setting, but this matter is left to future work.

[9] Thanks to Paul Christiano for this example.



(a) A one-step decision problem. (b) The task of maximizing a Q-function.

Fig. 6. Two possible representations of a Q-learner solving a one-step decision problem.

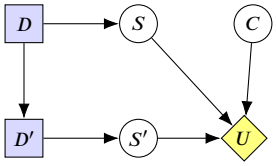


Fig. 7. The task of opening a combination lock.

in Fig. 6a, where D is chosen to optimise the reward R. Then, the future state s' satisfies the instrumental control incentive criterion. This matches our intuition — that RL systems may benefit from shaping their future environment. Suppose instead that we regard as an agent the function inside the Q-learner that chooses d to maximize the *Q-function* $q(s,d) := \mathbb{E}[R \mid s,d]$. In this model, shown in Fig. 6b.[10] Then, the decision's effect on s' is a mere side-effect to the task of maximizing $q(s,d)$. Although the instrumental control incentive is absent, the physical reality of this second scenario is identical to the first, and so there any harmful influence on s' may still be finely tuned to the agent's objective.

Ideally, we would reduce this sensitivity to modelling assumptions, and we might hope to achieve this by using more fundamental modelling assumptions, such as the independent causal mechanism assumption, to ascertain which variables should be viewed as decisions Kenton et al. [33, Sec. 4.3]. But such approaches still are sensitive to which variables are regarded as causal mechanisms or physical variables, and further research is needed to understand this dependence.

8. Incentives in a multi-decision setting

There are multiple possible ways that incentive concepts like RI, ICI and II may be generalized to multi-decision settings. This is because the presence of an incentive at some decision D may depend on the policy followed at other decisions. If we want to know the incentives when a model is fully trained, we could see whether some incentive concept ϕ holds for some or all of the optimal policies. Alternatively, we may be interested in sub-optimal policies as well. Both cases are included in the following definition. Note that in the case of a multi-decision CID, we will denote the decision rule for a particular decision D^i as π^i, and the set of decision rules for all other policies as $\boldsymbol{\pi}^{-i}$.[11]

Definition 14 *(Multi-decision ϕ-incentive). Let ϕ be a proposition defined on a single-decision SCIM, and let $\mathcal{M}$ be a multi-decision SCIM. There is an A- (resp. E-) optimal ϕ at the decision D^i if for all (resp. there exists some) $\boldsymbol{\pi} \in \arg\max_{\boldsymbol{\pi}'} \mathbb{E}_{\boldsymbol{\pi}'}[U]$, such that ϕ holds in $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$, the single-decision SCIM obtained by substituting in the decision rules $\boldsymbol{\pi}^{-i}$ for decisions other than D^i into $\mathcal{M}$.*

There is an A- (resp. E-) pre-optimal ϕ at D^i if for all (resp. there exists some) $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, such that ϕ holds in $\mathcal{M}_{\boldsymbol{\pi}^{-i}}$, where $\boldsymbol{\Pi}$ is the set of all policies.

We focus exclusively on cases where ϕ is the presence of a RI, II, or ICI, in a single-decision SCIM.

For example, consider the task of opening a combination lock (Fig. 7). Assume that the correct combination is $c = (9,9)$. There are two decisions, $d, d' \in \{0,\dots,9\}$, which are stored in the states $s = d$ and $s' = d'$, and that are checked against the combination to output utility of 1 or 0, i.e. $u = \delta(c[1] = s \wedge c[2] = s')$. If D is chosen optimally, i.e. $d = 9$, then S' has an instrumental control incentive for D', because it must be set to 9 in order to obtain $u = 1$. In other words, an A-optimal instrumental control incentive is present. If instead D is set to 8, then D' lacks any such incentive. So there is no A-pre-optimal instrumental control incentive.

[10] It would also be possible to consider a multi-agent influence diagram [27] where the Q function is included as a decision, and its goal is a loss function $\ell = |r - \hat{r}|$, but we note that the set of variables that satisfy the graphical criterion for an ICI would not be altered by including this Q variable, along with a utility variable ℓ that is a child of Q and R.

[11] Those familiar with temporal logic in games may notice that this is analogous to the notion of E-NASH and A-NASH propositions — ones that hold in one or all Nash Equilibria, respectively [10,64].

The fact that the instrumental control incentive is present for all optimal policies implies that it is also present for one optimal policy, i.e. that an E-optimal incentive is present, and for one policy altogether, i.e. that an E-pre-optimal incentive is also present. This is a general rule: the four types of multi-decision incentive always have this inclusion relation.

Proposition 2. *For any ϕ, A-pre-optimal incentive $\implies$ A-optimal incentive $\implies$ E-optimal incentive $\implies$ E-pre-optimal incentive.*

Proof. These implications, from left to right, hold because: i) Optimal policies are a subset of all policies, ii) any optimal policy is in the set of optimal policies, and iii) any optimal policy is a policy. □

In establishing graphical criteria for these incentive concepts, we can draw on a helpful equivalence. An E-pre-optimal incentive on D^i is equivalent to compatibility with a single-decision incentive on D^i, treating other decisions as chance variables. To see this, notice that in either case, one can impute any function to variables other than D^i. Since an E-pre-optimal incentive is the weakest of the four kinds of multi-decision incentive, the single-decision graphical criteria can be used to rule out *any* form of multi-decision incentive.

Proposition 3. *Let $\mathcal{M}$ be a multi-decision SCIM, and obtain $\mathcal{M}'$ by replacing all decisions except for D^i with chance nodes. If the graphical criterion for single-decision (RI/ICI/II) does not hold in $\mathcal{M}'$, then there is no A- or E-optimal or pre-optimal multi-decision (RI/ICI/II) in $\mathcal{M}$.*

Proof. Immediate from Proposition 2 and the fact that choosing a set of functions and distributions $\{f^j, P^j\}$ for D^{-i} such that $\mathcal{M}_{\{f^j,P^j\}_{j\neq i}}$ satisfies ϕ is equivalent to choosing a set of deterministic decision rules and distributions $\{\pi^j, P^j\}$ for D^{-i} such that $\mathcal{M}_{\{\pi^j,P^j\}_{j\neq i}}$ satisfies ϕ. □

9. Related work

Causal influence diagrams The use of structural functions in a causal influence diagram goes back to at least the *functional influence diagram* of Dawid [13]. The most similar alternative model is the Howard canonical form influence diagram [31,29]. 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 [28] and Shachter and Heckerman [54] only constrains the relationships to being partially causal, in that decisions are taken to be causally antecedent to their descendants (though adding new decision node parents to all nodes makes the diagram fully causal). Appendix A shows by example why the stronger causality property is necessary for most of the newly proposed incentive concepts. Building on this paper, multi-agent SCIMs are formalized in Hammond et al. [27], and an open-source Python implementation of CIDs has been developed [22].

Materiality and value of information The criterion for materiality, Theorem 9, builds on previous work. The concept of value of information was first introduced by Howard [30]. The materiality soundness proof follows previous proofs [56,41], while the completeness proof is most similar to an attempted proof by Nielsen and Jensen [48]. They propose the criterion $W \not\perp \boldsymbol{U}^D \mid \mathbf{Pa}_D$ for requisite nodes, which differs from (1) in the conditioned set. Taken literally,[12] their criterion is unsound for requisite nodes. For example, in Fig. 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.[13]

To have positive VoC, it is known that a node must be an ancestor of a utility node [55], but the authors know of no more specific criterion. The concept of a *relevant* node introduced by Nielsen and Jensen [48] also bears some resemblance to VoC.

The relation of the current technical results to prior work is summarised in Table 1.

Instrumental control incentives and intent In a causal setting, Kleiman-Weiner et al. [35] offered a notion of intention to influence a variable O. A different kind of approach was taken by Halpern and Kleiman-Weiner [26] and Ward et al. [63], which offered definitions of intent that are specific to outcomes $O = o$. In particular, Ward et al. [63] was the first to prove a graphical criterion for any version of intent. We extend this work by defining a positive version of intent, rather than just considering negative intent, and by proving a graphical criterion for this new concept.

AI fairness Another application of this work is to evaluate when an AI system is incentivized to behave unfairly, on some definition of fairness. Response incentives address this question for counterfactual fairness [38,34]. An incentive criterion corresponding to path-specific effects [66,46] has been established by Ashurst et al. [5], for the single-decision setting. Nabi et al. [47] have shown

[12] Def. 3 defines d-separation for potentially overlapping sets.

[13] Furthermore, to prove that nodes meeting the d-connectedness property are requisite, Nielsen and Jensen claim that "X is [requisite] for D if $\mathrm{P}(\mathrm{dom}(U) \mid D, \mathbf{Pa}^D)$ 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. Since an earlier version of this paper was placed online [17], this completeness result was independently discovered by Zhang et al. [65, Thm. 2] and Lee and Bareinboim [42, Thm. 1]. There has also been further work in generalising this result to the case of multi-decision influence diagrams, in Van Merwijk et al. [61], where a sound and complete criterion is known for a class of influence diagrams said to satisfy "solubility", also known as "sufficient recall".

Table 1
Comparison with previous work, in a single-decision setting. The concept of materiality is well-known. For VoI, a new, corrected proof is provided. For VoC, the present work offers a new criterion, proving it sound and complete. For response incentive (RI) and instrumental control incentive (ICI), the criterion and all proofs are new.

	Definition	Criterion	Soundness	Completeness
Mater-iality	Howard [30]; Matheson [43]	Fagiuoli and Zaffalon [20]; Lauritzen and Nilsson [41]; Shachter [57]	Fagiuoli and Zaffalon [20]; Lauritzen and Nilsson [41]; Shachter [57]	First correct proof to our knowledge; see section 9
RI	New	New	New; proved using do-calculus and d-sep	New; proved constructively
ICI	New	New	New; proved using do-calculus	New; proved constructively
(Positive/ negative) intent	(Halpern and Kleiman-Weiner [26]/new)	(Ward et al. [63]/ new)	(Ward et al. [63]/ new)	(Ward et al. [63]/ new)
II	New	New	New; proved using do-calculus	New; proved constructively

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.

Rational verification Verification is the task of checking that a program satisfies specified properties, which is relevant to the present study because we are proposing to use incentive concepts to check agent behaviour. Typically, specifications are defined using temporal logic [15]; sometimes a probabilistic temporal logic is used [39]. Of particular relevance is "rational verification", which validates the behaviour of agents that are pursuing objectives [1,24,64]. Overall, our work aligns with rational verification in that it aims to verify agent behaviour. The difference is that we have explored what kinds of properties can be specified in the language of causality in particular (rather than, for example, a temporal logic). Relatedly, rather than using a Kripke structure of partially observable Markov decision process to model an agent-environment interaction, we have used causal models.

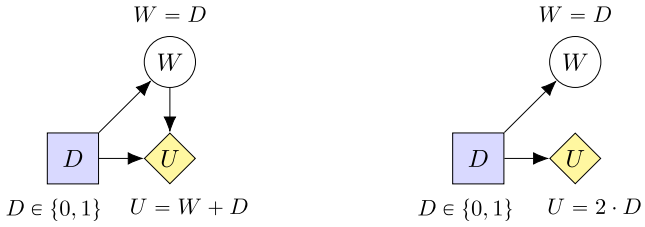
Mechanism design The aim of mechanism design is to understand how objectives and environments can be designed, in order to shape the behaviour of rational agents (e.g. 49, Part II). At this high level, mechanism design is closely related to the incentive design results we have developed in this paper. In practice, however, the strands of research look rather different. Whereas mechanism design is primarily concerned with defining objective functions and action spaces that ensure desirable Nash equilibria, our core interest is on defining specifications for safe and fair agent behaviour, and on the causal structures that ensure that these specifications are satisfied.

10. Discussion and conclusion

We have defined three new concepts: response incentives, instrumental control incentives and impact incentives, and have spelled out the connection between ICIs and an existing concept, intent. We have proved complete graphical criteria for all four concepts in a single-decision setting. Moreover, we have introduced a notion of incentives for influence diagrams with multiple decisions, and proved that the criteria are also sound for those cases. In all cases we have shown how these definitions have implications for other concepts of broader interest, such as instrumental goals, counterfactual fairness, and impact measures. We have also shown via toy examples how different existing approaches might be appropriate to addressing different kinds of problems, and have outlined circumstances in which each kind of approach is favoured. These incentive concepts have already seen applications to areas including value learning [4], interruptibility [40], conservatism [11], modelling agent frameworks [16] and reward tampering [19].

Let us now outline some limitations of this paper, and what they might mean for future work. First, note that to apply these criteria, we require knowledge of the (causal) structure of the interaction between agent and environment. Sometimes, experts know these causal relationships even when they do not know the exact parametric relationships between variables — an ideal use case for these criteria. In the context of incentive design, such a scenario may often arise, since these causal relationships often follow directly from the design choices for an agent and its objective. Sometimes, however, we may have too little knowledge of the causal structure to be able to apply the criteria. In other cases, we may have, in a sense, too much knowledge for the graphical criteria to be useful. With abundant experimental data, we might compute safety and fairness properties (such as counterfactual fairness) directly, removing any need for the incentive concepts and graphical criteria. A fourth scenario is that the world is not even describable by a fixed graphical model, but rather it is better understood using a probability tree, or relatedly, as an extensive form game. These limitations suggest possible avenues for future work. To enlarge the set of cases in which incentives can be evaluated, it may be possible to devise ways of combining experimental data with a priori knowledge to arrive at an evaluation. To deal with extensive form games, it may be possible to devise graphical criteria for probability tree and game trees.

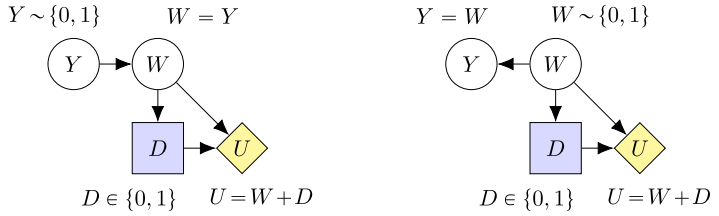
Another limitation of graphical criteria is that they can only offer a definitive resolution in one direction. Also, although they can rule out incentives definitively, they can only rule that the presence of an incentive is compatible with the graphical structure. It is still yet to be established how often incentives are present when they are compatible with the graph. This might be proved using measure theoretic arguments resembling the arguments that d-connection almost always implies conditional dependence [44].



(a) A causal influence diagram reflecting the causal structure of the environment

(b) Influence diagram that is causal in the sense of Heckerman and Shachter [28, 29]

Fig. A.8. Two different influence diagram representations of the same situation, with different VoC and ICI.



(a) A causal influence diagram reflecting the causal structure of the environment

(b) Influence diagram that is causal in the sense of Heckerman and Shachter [28, 29]

Fig. A.9. Two different influence diagram representations of the same situation, with different RI and VoC. In Fig. A.9a, Y is sampled from some arbitrary distribution on $\{0,1\}$, for example a Bernoulli distribution with $p=0.5$. In Fig. A.9b, W is sampled in the same way.

Relatedly, their output says nothing of the strength of incentive present, which can only be established using detailed knowledge of the strength of causal relationships present in the environment, rather than just their presence or absence.

Finally, it would be possible to improve the applicability of these graphical criteria by extending them to multi-agent settings. So far, we have considered single-agent settings, where the world is divided into agent and environment. If instead part of the environment was modelled as a rival agent, and we assume Nash Equilibrium policy profiles, then this would place additional constraints on how that part of the environment may behave. So, in some cases where single-agent criteria cannot rule out an incentive, a multi-agent criterion should be able to rule out that incentive. On the other hand, if it is known that another player will observe and respond strategically to one's policy, then this could mean that policies could influence one another via pathways that are not visible in the original causal graph, which could mean that multi-agent incentives might arise, when the criteria for a single-agent setting would have ruled them impossible. Some groundwork has been laid by [27], which formalizes multi-agent influence diagrams, but a full analysis of the multi-agent setting is left to future work.

CRediT authorship contribution statement

Ryan Carey: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Eric Langlois:** Formal analysis, Conceptualization. **Chris van Merwijk:** Formal analysis, Conceptualization. **Shane Legg:** Supervision. **Tom Everitt:** Supervision, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Causality examples

Causal influence diagrams that reflect the full causal structure of the environment are needed to correctly capture response incentives, value of control and instrumental control incentives. We begin with showing this for instrumental control incentives and value of control, leaving response incentive to the end of this section. Consider the two influence diagrams in Fig. A.8. If we assume that W really affects U, only the diagram in Fig. A.8a correctly represents this causal structure, whereas Fig. A.8b lacks the edge $W \to U$. According to Definitions 11 and 16, W has positive value of control and an instrumental control incentive. Only Fig. A.8a gets this right.

The influence diagram literature has discussed weaker notions of causality, under which Fig. A.8b is considered a valid alternative representation of the situation described by Fig. A.8a. For example, if we only consider their joint distributions conditional on various policies, then Figs. A.8a and A.8b are identical. Both diagrams are also in the canonical form of Heckerman and Shachter [29], as

every variable responsive to the decision is a descendant of the decision. For the same reason, both diagrams are also causal influence diagrams in the terminology of Heckerman and Shachter [28] and Shachter and Heckerman [54]. Since only Fig. A.8a gets the incentives right, we see that the stronger notion of causal influence diagram introduced in this paper is necessary to correctly model instrumental control incentives and value of control.

To show that response incentives also rely on fully causal influence diagrams, consider the diagrams in Fig. A.9. Again, we assume that Fig. A.9a accurately depicts the environment, while Fig. A.9b has the edge $Y \to W$ reversed. Again, both diagrams have identical joint distributions given any policy. Both diagrams are also causal in the weaker sense of Heckerman and Shachter [28] and Shachter and Heckerman [54]. Yet only the fully causal influence diagram in Fig. A.9a exhibits that Y can have a response incentive or positive value of control.

Appendix B. 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 value of information [30,43]. To assess VoI for variables $\boldsymbol{W}$, we first make $\boldsymbol{W}$ an observation by adding a link $W \to D$ for each $W \in \boldsymbol{W}$ and then test whether any W is material in the updated model [57].

Definition 15 *(Value of information). Nodes $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \boldsymbol{Desc}^D$ in a single-decision SCIM $\mathcal{M}$ have* VoI *if $\mathcal{V}^*(\mathcal{M}_{\boldsymbol{W} \not\to D}) < \mathcal{V}^*(\mathcal{M}_{\boldsymbol{W} \to D})$ where $\mathcal{M}_{\boldsymbol{W} \to D}$ is obtained from $\mathcal{M}$ by adding the edges from each $W \in \boldsymbol{W}$ to D, and $\mathcal{M}_{\boldsymbol{W} \not\to D}$ is obtained by removing them.*

Since Definition 15 adds an information link, it can only be applied to variables $\boldsymbol{W}$ that are non-descendants of the decision, lest cycles be created in the graph.

We will say that a CID $\mathcal{G}$ *admits* VoI for $\boldsymbol{W}$ if $\boldsymbol{W}$ has VoI in a SCIM $\mathcal{M}$ compatible with $\mathcal{G}$. More generally, for any proposition ϕ, we will say that $\mathcal{G}$ admits ϕ if there exists any SCIM $\mathcal{M}$ compatible with $\mathcal{G}$ that satisfies ϕ.

An observed variable having positive VoI means that it would be material if it was observed. Using this insight, we can adapt the criterion from Definition 7 to check for positive VoI. For a latent variable, we add an edge from it to the decision, and then check the graphical criterion. We prove that this procedure is tight, in that it identifies every zero VoI node that can be identified from the graphical structure (in a single decision setting).

Theorem 8 *(Value of information criterion). A single decision CID $\mathcal{G}$ admits VoI for $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \boldsymbol{Desc}^D$ if and only if there exists some $W \in \boldsymbol{W}$ that is a requisite observation in $\mathcal{G}_{\boldsymbol{W} \to D}$, the graph obtained by adding edges from $\boldsymbol{W}$ to D, to $\mathcal{G}$.*

The proof is deferred to Appendix E.5.

Appendix C. Value of control

So far, we have considered what information an agent would like to know, or be influenced by. We now consider what variables an agent would like to control. A variable has VoC if a decision-maker could benefit from setting its value [55,43,54]. Concretely, we ask whether the attainable utility can be increased by letting the agent decide the structural function for the variable.

Definition 16 *(Value of control). In a single-decision SCIM $\mathcal{M}$, the set of non-decision nodes $\boldsymbol{W}$ has* positive value of control *if*

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

where $g^{\boldsymbol{W}}$ is a set of soft interventions for $\boldsymbol{W}$, i.e. a new structural function $g^W : \mathrm{dom}(\mathbf{Pa}^W \cup \{\mathcal{E}^W\}) \to \mathrm{dom}(W)$ that respects the graph, for each $W \in \boldsymbol{W}$.

This can be deduced from the graph, using again the minimal reduction (Definition 9) to rule out effects through observations that an optimal policy can ignore.

Theorem 9 *(Value of control criterion). A single-decision CID $\mathcal{G}$ admits positive value of control for non-decision vertices $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \{D\}$ if and only if there is a directed path $W \dashrightarrow U$ for some $W \in \boldsymbol{W}$ and $U \in \boldsymbol{U}$ in the minimal reduction $\mathcal{G}^{min}$.*

The proof is supplied in Appendix E.9.

To apply this criterion to the content recommendation example (Fig. 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.

Appendix D. Intent equivalence

First, let us restate our definition.

Definition 12 *(Intent). Let $\mathcal{M}$ be a single-decision SCIM that represents an agent's beliefs. There is* additive intent *to influence nodes $\boldsymbol{W}$ by choosing $\boldsymbol{\pi}^*$ over $\boldsymbol{\pi}'$ if $\mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] < \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}]$, and $\boldsymbol{W}$ is a subset $\boldsymbol{W} \subseteq \boldsymbol{Z}$ of variables $\boldsymbol{Z}$, that is subset-minimal such that:*

$$\mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}_{\boldsymbol{Z}_{\boldsymbol{\pi}^*}}] \geq \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}]. \tag{3}$$

There is subtractive intent *if $\mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] < \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}]$ and $\boldsymbol{Z}$ is subset-minimal such that:*

$$\mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}_{\boldsymbol{Z}_{\boldsymbol{\pi}'}}] \leq \mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}]. \tag{4}$$

For a set $\boldsymbol{\Pi}'$, we say that there is an (additive/subtractive) intent to influence $\boldsymbol{W}$ by choosing $\boldsymbol{\pi}^*$ over $\boldsymbol{\pi}'$ if this intent is present over every $\boldsymbol{\pi}'$ in $\boldsymbol{\Pi}'$.

And here is Halpern's definition, translated into an SCIM setting.

Definition 17 *(Intent; adapted from Def. 4.4 of [26]). In a single-decision SCIM $\mathcal{M}$, an agent intends to affect $\boldsymbol{W}$ by choosing policy $\boldsymbol{\pi}$ and reference set $\boldsymbol{\Pi}'$ if there exists a superset $\boldsymbol{Z} \supseteq \boldsymbol{W}$ such that: a) $\mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}}] < \max_{\boldsymbol{\pi}'} \mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}', \boldsymbol{Z}_{\boldsymbol{\pi}}}]$, and b) $\boldsymbol{Z}$ is subset-minimal; i.e. for any strict subset $\boldsymbol{Z}^*$, we have $\mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}}] \geq \max_{\boldsymbol{\pi}'} \mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}', \boldsymbol{Z}^*_{\boldsymbol{\pi}}}]$.*

We now prove that for a non-empty set $\boldsymbol{W}$ of variables, Halpern's definition matches our own.

Theorem 10. *For a non-empty set of variables $\boldsymbol{W}$, the presence of additive Intent is equivalent to an agent intending to affect $\boldsymbol{W}$ in Halpern's definition.*

Proof. *Proof that subtractive intent implies Halpern intent* If there is additive intent over every $\boldsymbol{\pi}' \in \boldsymbol{\Pi}'$, then $\mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}}] < \mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}', \boldsymbol{Z}_{\boldsymbol{\pi}}}]$ for every $\boldsymbol{\pi} \in \boldsymbol{\Pi}'$, and so $\mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}}] < \max_{\boldsymbol{\pi}'} \mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}', \boldsymbol{Z}_{\boldsymbol{\pi}}}]$, implying Halpern intent. *Proof that Halpern intent implies additive intent* To begin with, if $\mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] \geq \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}]$, then we would have that $\boldsymbol{Z} = \emptyset$ would always satisfy (a), and so there could not exist any non-empty set $\boldsymbol{W}$ satisfying Halpern intent. Since $\boldsymbol{W}$ is assumed to be non-empty, we must therefore have $\mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] < \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}]$, satisfying the first condition of additive intent. Moreover, if $\mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}}] < \max_{\boldsymbol{\pi}'} \mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}', \boldsymbol{Z}_{\boldsymbol{\pi}}}]$ we have $\mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}}] < \mathbb{E}[\mathcal{U}_{\boldsymbol{\pi}', \boldsymbol{Z}_{\boldsymbol{\pi}}}]$ for every $\boldsymbol{\pi} \in \boldsymbol{\Pi}'$, satisfying the other condition, meaning that additive intent is present. □

Appendix E. Proofs

E.1. Preliminaries

Our proofs will rely on the following fundamental results about causal models from [12], [23] and [52].

Definition 18 *(Causal Irrelevance). $\boldsymbol{X}$ is* causally irrelevant *to $\boldsymbol{Y}$, given $\boldsymbol{Z}$, written $(\boldsymbol{X} \nrightarrow \boldsymbol{Y} | \boldsymbol{Z})$ if, for every set $\boldsymbol{W}$ disjoint of $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$, we have*

$$\forall \varepsilon, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{x}', \boldsymbol{w} \qquad \boldsymbol{Y}_{\boldsymbol{xzw}}(\varepsilon) = \boldsymbol{Y}_{\boldsymbol{x'zw}}(\varepsilon)$$

Lemma 1. *Recall that $(\boldsymbol{X} \not\rightarrow \boldsymbol{Y} | \boldsymbol{Z})_{\mathcal{G}}$ means that $\mathcal{G}$ contains no directed path from $\boldsymbol{X}$ to $\boldsymbol{Y}$, except possibly through $\boldsymbol{Z}$. Then, for every SCM $\mathcal{M}$ compatible with a DAG $\mathcal{G}$,*

$$(\boldsymbol{X} \not\rightarrow \boldsymbol{Y} | \boldsymbol{Z})_{\mathcal{G}} \Rightarrow (\boldsymbol{X} \nrightarrow \boldsymbol{Y} | \boldsymbol{Z})$$

Proof. By induction over variables, as in [23, Lemma 12]. □

Lemma 2 *(52, Thm. 3.4.1, Rule 1). For any disjoint subsets of variables $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ in the DAG $\mathcal{G}$, $\mathbb{E}(\boldsymbol{Y}_{\boldsymbol{x}} | \boldsymbol{z}, \boldsymbol{w}) = \mathbb{E}(\boldsymbol{Y}_{\boldsymbol{x}} | \boldsymbol{w})$ if $\boldsymbol{Y} \perp \boldsymbol{Z} | (\boldsymbol{X}, \boldsymbol{W})$ in the graph $\mathcal{G}'$ formed by deleting all incoming edges to $\boldsymbol{X}$.*

Lemma 3 *(52, Thm. 1.2.4). For any three disjoint subsets of nodes $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ in a DAG $\mathcal{G}$, $(\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} | \boldsymbol{Z})$ if and only if $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z})_P$ for every probability function P compatible with $\mathcal{G}$.*

Lemma 4 *(12, Sigma Calculus Rule 3). For any disjoint subsets of nodes $(\boldsymbol{X}, \boldsymbol{Y}) \subseteq \boldsymbol{V}$ and $\boldsymbol{Z} \subseteq \boldsymbol{V}$ in a DAG $\mathcal{G}$ $\mathrm{P}(\boldsymbol{X} | \boldsymbol{Z}; g^{\boldsymbol{Y}}) = \mathrm{P}(\boldsymbol{X} | \boldsymbol{Z}; g'^{\boldsymbol{Y}})$ if $\boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z}$ in $\mathcal{G}_{\overline{\boldsymbol{Y}(\boldsymbol{Z})}}$ where $\boldsymbol{Y}(\boldsymbol{Z}) \subseteq \boldsymbol{Y}$ is the set of elements in $\boldsymbol{Y}$ that are not ancestors of $\boldsymbol{Z}$ in $\mathcal{G}$ and $\mathcal{G}_{\overline{\boldsymbol{W}}}$ denotes $\mathcal{G}$ but with edges incoming to variables in $\boldsymbol{W}$ removed.*

E.2. An optimal policy that respects the minimal reduction

First, we introduce the notion of a $\mathcal{G}^{\min}$-respecting optimal policy. Our proof of its optimality is similar to Theorem 3 from [41]. It builds on the following intersection property of d-separation.

Lemma 5 *(d-separation* intersection *property). For all disjoint sets of variables $\boldsymbol{W}$, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$,*

$$(\boldsymbol{W} \perp \boldsymbol{X} | \boldsymbol{Y}, \boldsymbol{Z}) \wedge (\boldsymbol{W} \perp \boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{Z}) \Rightarrow (\boldsymbol{W} \perp (\boldsymbol{X} \cup \boldsymbol{Y}) | \boldsymbol{Z})$$

Proof. Suppose that the RHS is false, so there is a path from $\boldsymbol{W}$ to $\boldsymbol{X} \cup \boldsymbol{Y}$ conditional on $\boldsymbol{Z}$. This path must have a sub-path that passes from $\boldsymbol{W}$ to $X \in \boldsymbol{X}$ without passing through $\boldsymbol{Y}$ or to $Y \in \boldsymbol{Y}$ without passing through $\boldsymbol{X}$ (it must traverse one set first). But this implies that $\boldsymbol{W}$ is d-connected to $\boldsymbol{X}$ given $\boldsymbol{Y}, \boldsymbol{Z}$ or to $\boldsymbol{Y}$ given $\boldsymbol{X}, \boldsymbol{Z}$, meaning the LHS is false. So if the LHS is true, then the RHS must be true. □

Lemma 6 *($\mathcal{G}^{min}$-respecting optimal policy). Every single-decision SCIM $\mathcal{M} = \langle \mathcal{E}, \boldsymbol{V}, \boldsymbol{F}, P, \boldsymbol{U}, \mathcal{O} \rangle$ has an optimal policy $\tilde{\pi}$ that depends only on requisite observations. In other words, $\tilde{\pi}$ is also a policy for the minimal model $\mathcal{M}^{min} = \left\langle \mathcal{G}^{min}, \mathcal{E}, \boldsymbol{F}, P \right\rangle$. We call $\tilde{\pi}$ a $\mathcal{G}^{\min}$-respecting* optimal policy.

This result is already known from [41,20], but we prove it here to make the paper more self-contained.

Proof. First partition $\mathbf{Pa}_{\mathcal{G}}^{D}$ into the requisite parents $\mathbf{Pa}_{\min}^{D} = \{W \in \mathbf{Pa}^{D} : W \not\perp \mathcal{G}\boldsymbol{U}^{D} \mid \{D\} \cup \mathbf{Pa}^{D} \setminus \{W\}\}$, and non-requisite parents $\mathbf{Pa}_{\_}^{D} = \mathbf{Pa}_{\mathcal{G}}^{D} \setminus \mathbf{Pa}_{\min}^{D}$.

Let π^* be an optimal policy in $\mathcal{M}$. To construct a $\mathcal{G}^{\min}$-respecting version $\tilde{\pi}$, select any value $\tilde{\mathbf{pa}}_{\_}^{D} \in \mathrm{dom}(\mathbf{Pa}_{\_}^{D})$ for which $\mathrm{P}_{\pi^*}(\mathbf{Pa}_{\_}^{D} = \tilde{\mathbf{pa}}_{\_}^{D}) > 0$. For all $\mathbf{pa}_{\min}^{D} \in \mathrm{dom}(\mathbf{Pa}_{\min}^{D})$ and $\varepsilon^{D} \in \mathrm{dom}(\mathcal{E}^{D})$, let

$$\tilde{\pi}(\mathbf{pa}_{\min}^{D}, \mathbf{pa}_{\_}^{D}, \varepsilon^{D}) := \pi^*(\mathbf{pa}_{\min}^{D}, \tilde{\mathbf{pa}}_{\_}^{D}, \varepsilon^{D}).$$

The policy $\tilde{\pi}$ is permitted in $\mathcal{M}^{\min}$ because it does not vary with $\mathbf{Pa}_{\_}^{D}$.

Now let us prove that $\tilde{\pi}$ is optimal in $\mathcal{M}$. Partition U into $\boldsymbol{U}^{D} = \boldsymbol{U} \cap \mathbf{Desc}^{D}$ and $\boldsymbol{U}^{\setminus D} = \boldsymbol{U} \setminus \mathbf{Desc}^{D}$. D is causally irrelevant for every $U \in \boldsymbol{U}^{\setminus D}$ so every policy π (in particular, $\tilde{\pi}$) is optimal with respect to $\mathcal{U}^{\setminus D} := \sum_{U \in \boldsymbol{U}^{\setminus D}} U$.

We now consider $\boldsymbol{U}^{D}$. By definition, $W \perp \boldsymbol{U}^{D} \mid \{D\} \cup \mathbf{Pa}^{D} \setminus \{W\}$ for every $W \in \mathbf{Pa}_{\_}^{D}$. By inductively applying the intersection property of d-separation (Lemma 5) over elements of $\mathbf{Pa}_{\_}^{D}$ we obtain

$$\mathbf{Pa}_{\_}^{D} \perp \boldsymbol{U}^{D} \mid \{D\} \cup \mathbf{Pa}_{\min}^{D}. \tag{E.1}$$

Next, we establish that $\mathbb{E}_{\tilde{\pi}}[\mathcal{U}^{D}] = \mathbb{E}_{\pi^*}[\mathcal{U}^{D}]$ by showing that $\mathbb{E}_{\tilde{\pi}}[\mathcal{U}^{D} \mid \mathbf{pa}^{D}] = \mathbb{E}_{\pi^*}[\mathcal{U}^{D} \mid \mathbf{pa}^{D}]$ for every $\mathbf{pa}^{D} \in \mathrm{dom}(\mathbf{Pa}^{D})$ with $\mathrm{P}(\mathbf{pa}^{D}) > 0$. First, the expected utility of $\tilde{\pi}$ given any $(\mathbf{pa}_{\min}^{D}, \mathbf{pa}_{\_}^{D})$ with $\mathrm{P}(\mathbf{Pa}_{\min}^{D} = \mathbf{pa}_{\min}^{D}, \mathbf{pa}_{\_}^{D} = \mathbf{pa}_{\_}^{D}) > 0$ is equal to the expected utility of π^* on input $(\mathbf{pa}_{\min}^{D}, \tilde{\mathbf{pa}}_{\_}^{D})$:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}}[\mathcal{U}^{D} \mid \mathbf{pa}_{\min}^{D}, \mathbf{pa}_{_}^{D}] \\
&\quad = \sum_{u,d} \Big(u\, \mathrm{P}(\mathcal{U}^{D} = u \mid d, \mathbf{pa}_{\min}^{D}, \mathbf{pa}_{_}^{D}) \\
&\qquad\qquad \cdot \mathrm{P}_{\tilde{\pi}}(D = d \mid \mathbf{pa}_{\min}^{D}, \mathbf{pa}_{_}^{D}) \Big) \\
&\quad = \sum_{u,d} \Big(u\, \mathrm{P}(\mathcal{U}^{D} = u \mid d, \mathbf{pa}_{\min}^{D}, \tilde{\mathbf{pa}}_{_}^{D}) \\
&\qquad\qquad \cdot \mathrm{P}_{\pi^*}(D = d \mid \mathbf{pa}_{\min}^{D}, \tilde{\mathbf{pa}}_{_}^{D}) \Big) \\
&\quad = \mathbb{E}_{\pi^*}[\mathcal{U}^{D} \mid \mathbf{pa}_{\min}^{D}, \tilde{\mathbf{pa}}_{_}^{D}]
\end{aligned}$$

where the middle equality follows from (E.1) and the definition of $\tilde{\pi}$. Second, the expected utility of π^* given input $\tilde{\mathbf{pa}}_{\_}^{D}$ is the same as its expected utility on any input $\mathbf{pa}_{\_}^{D}$:

$$\begin{aligned}
&= \max_{d} \mathbb{E}_{\pi^*}[\mathcal{U}_{d}^{D} \mid \mathbf{pa}_{\min}^{D}, \tilde{\mathbf{pa}}_{_}^{D}] \\
&= \max_{d} \mathbb{E}_{\pi^*}[\mathcal{U}_{d}^{D} \mid \mathbf{pa}_{\min}^{D}, \mathbf{pa}_{_}^{D}] \\
&= \mathbb{E}_{\pi^*}[\mathcal{U}^{D} \mid \mathbf{pa}_{\min}^{D}, \mathbf{pa}_{_}^{D}]
\end{aligned}$$

where the first equality follows from the optimality of π^* and the second from Lemma 2. The expression $\mathbb{E}_{\pi^*}[\mathcal{U}_{d}^{D} \mid \cdots]$ means that we first assign the policy π^* then intervene to set $D = d$, which renders π^* effectively irrelevant but formally necessary for creating an SCM. This result shows that $\tilde{\pi}$ is optimal for $\mathcal{U}^{D}$ and has $\mathbb{E}_{\tilde{\pi}}[\mathcal{U}^{D}] = \mathbb{E}_{\pi^*}[\mathcal{U}^{D}]$. Since $\tilde{\pi}$ is optimal for both $\mathcal{U}^{D}$ and $\mathcal{U}^{\setminus D}$, $\tilde{\pi}$ is optimal in $\mathcal{M}$. □

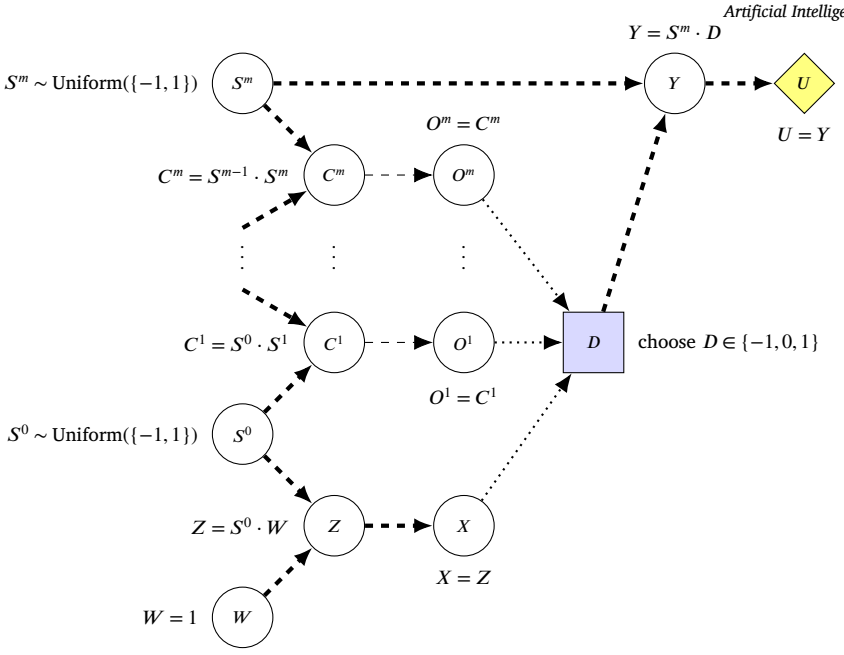


Fig. E.10. Outline of the variables involved in the response incentive construction. Every graph that satisfies the response incentive graphical criterion contains this structure (allowing all dashed paths except those to C^i or Y to have length zero). An optimal policy for the given model is $D = X \cdot \prod_i O^i = S^m$, yielding utility $U = Y = W(S^m)^2 = 1$, and all optimal policies must depend on the value of X.

E.3. Response incentive criterion

We now prove the soundness and completeness of the response incentive criterion.

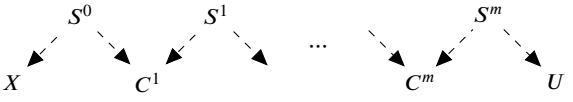
Theorem 3 *(Response incentive criterion). A single-decision CID $\mathcal{G}$ admits a response incentive on $\boldsymbol{W} \subseteq \boldsymbol{X}$ if and only if the minimal reduction $\mathcal{G}^{min}$ has a directed path $W \dashrightarrow D$ for some $W \in \boldsymbol{W}$.*

Proof of Theorem 3. We first prove that the criterion is sound, and then that it is complete.

Soundness (the *only if* direction). For the soundness direction, assume that for $\mathcal{G}$, the minimal reduction $\mathcal{G}^{\min}$ contains no directed path $W \dashrightarrow D$ for any $W \in \boldsymbol{W}$. Let $\mathcal{M} = \langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}, P, \boldsymbol{U}, \mathcal{O} \rangle$ be any SCIM compatible with $\mathcal{G}$. Let $\mathcal{M}^{\min} = \left\langle \mathcal{G}^{\min}, \boldsymbol{\mathcal{E}}, \boldsymbol{F}, P \right\rangle$ be $\mathcal{M}$, but with the minimal reduction $\mathcal{G}^{\min}$. By Lemma 6 in appendix E, there exists a $\mathcal{G}^{\min}$-respecting policy $\tilde{\pi}$ that is optimal in $\mathcal{M}$. In $\mathcal{M}^{\min}_{\tilde{\pi}}$, $\boldsymbol{W}$ is causally irrelevant for D, so $D(\boldsymbol{\varepsilon}) = D_{g_{\boldsymbol{W}}}(\boldsymbol{\varepsilon})$. Furthermore, $\mathcal{M}_{\tilde{\pi}}$ and $\mathcal{M}^{\min}_{\tilde{\pi}}$ are the same SCM, with the functions $\boldsymbol{F} \cup \{\tilde{\pi}\}$. So $D(\boldsymbol{\varepsilon}) = D_{g_{\boldsymbol{W}}}(\boldsymbol{\varepsilon})$ also in $\mathcal{M}_{\tilde{\pi}}$, which means that there is an optimal policy in $\mathcal{M}$ that does not respond to interventions on $\boldsymbol{W}$ for any $\boldsymbol{\varepsilon}$.

Completeness (the *if* direction). Fig. E.10 illustrates the model constructed in the proof.

Starting from the assumption that there exists $W \in \boldsymbol{W}$ with $W \dashrightarrow D$ in $\mathcal{G}^{\min}$, we explicitly construct a compatible model for $\mathcal{G}$ for which the decision of every optimal policy causally depends on the value of W. Let $\overrightarrow{WD}$ be a directed path from W to D that only contains a single requisite observation that we label X (if W is itself a requisite observation, then W and X are the same node). Since X is a requisite observation for D, there exists some utility node U descending from D that is d-connected to X in $\mathcal{G}$ when conditioning on $\mathbf{Pa}^D \cup \{D\} \setminus \{X\}$. Let $\overrightarrow{DU}$ be a directed path from D to U and let $\overrightarrow{XU}$ be a path between X and U that is active when conditioning on $\mathbf{Pa}^D \cup \{D\} \setminus \{X\}$. By the definition of d-connecting paths, $\overrightarrow{XU}$ has the following structure ($m \geq 0$):



consisting of directed sub-paths leaving source nodes S^i and entering collider nodes C^i, where there is a directed path from each collider to $\mathbf{Pa}^D \cup \{D\} \setminus \{X\}$ and no non-collider node is in $\mathbf{Pa}^D \cup \{D\} \setminus \{X\}$. It may be the case that X and S^0 are the same node. For each $i \in \{1, \ldots, m\}$, let $\overrightarrow{C^i O^i}$ be a directed path from C^i to some $O^i \in \mathbf{Pa}^D$ such that no other node along $\overrightarrow{C^i O^i}$ is in $\mathbf{Pa}^D$.

We make the following assumptions without loss of generality:

- $\overrightarrow{XU}$ first intersects $\overrightarrow{DU}$ at some variable Y (possibly Y is U) and thereafter both $\overrightarrow{XU}$ and $\overrightarrow{DU}$ follow the same directed path from Y to U (otherwise, let Y be the first intersection point and replace the $Y \dashrightarrow U$ sub-path of $\overrightarrow{XU}$ with the $Y \dashrightarrow U$ sub-path of $\overrightarrow{DU}$).

- The $S^0 \dashrightarrow X$ sub-path of reversed $\overline{XU}$ first intersects $\overrightarrow{WD}$ at some node Z and thereafter both follow the same directed path from Z to X (same argument as for Y).
- The paths $\overrightarrow{C^iO^i}$ are mutually non-intersecting (if there is an intersection between $\overrightarrow{C^iO^i}$ and $\overrightarrow{C^jO^j}$ with $j \neq i$ then replace the part of $\overline{XU}$ between C^i and C^j with the path through the intersection point, which becomes the new collider; this can only happen finitely many times as it reduces the number of collider nodes).

The resulting structure is shown in Fig. E.10.

We now formally define the model represented in the figure. The domains of all endogenous variables are set to $\{-1, 0, 1\}$. All exogenous variables are given independent discrete uniform distributions over $\{-1, 1\}$. Unless otherwise specified, we set $B = A$ for each edge $A \to B$ within the directed paths shown in Fig. E.10, i.e. $f^B(\mathbf{pa}^B, \varepsilon^B) = a$. Nodes at the heads of directed paths can therefore be defined in terms of nodes at the tails. We begin by describing functions for the "default" case depicted by Fig. E.10, and discuss adaptations for various special cases below.

- $S^i = \mathcal{E}^{S^i}$, giving S^i a uniform distribution over -1 and 1.
- $U = Y$, and
- $Y = S^m \cdot D$, so D must match S^m to optimize utility.
- $C^i = S^{i-1} \cdot S^i$, and
- $O^i = C^i$, so the collider C^i reveals (only) whether S^{i-1} and S^i have the same sign or not.
- $W = 1$,
- $Z = W \cdot S^0$, and
- $X = Z$, so X reflects the value of S^0, unless W is intervened upon.

All other variables not part of any named path are set to 0.

Special cases arise when two or more of the labeled nodes in Fig. E.10 refer to the same variable. When X, Y, or O^i is the same node as one of its parents, then it simply takes the function of this parent (instead of copying its value). Meanwhile, the S^i, C^i, and Y nodes must be distinct by construction, so no special cases treatment is required. Finally, the functions for W, S^0 and Z are adapted per the following cases:

Case 1: W, S^0, and Z are all the same node. Let $W = Z = S^0 = \mathcal{E}^{S^0}$, i.e. the node takes a uniform distribution over $\{-1, 1\}$.

Case 2: Z is the same node as S^0, but different from W. In this case, let $Z = S^0 = W \cdot \mathcal{E}^{S^0}$.

Case 3: W is the same node as Z, but different from S^0. In this case, let $W = Z = S^0$.

The final combination of W and S^0 being the same, while different from Z, cannot happen by the definition of Z.

Regardless of which case applies, an optimal policy is $D = X \cdot \prod_{i=1}^m O^i$, which yields a utility of 1.

Let $g^{\boldsymbol{W}}$ be the intervention $\mathtt{do}(W = 0)$. Formally, $g^{\boldsymbol{W}}$ has g^W deterministically set $W = 0$, and applies the unchanged function $g^{W'} = f^{W'}$ for the other variables $W' \in \boldsymbol{W} \setminus \{W\}$. Under $g^{\boldsymbol{W}}$, it follows that $X_{W=0} = Z_{W=0} = 0$. Without the information in X, S^m is independent of $(\mathbf{Pa}^D)_{W=0}$ and hence is independent of $D_{W=0}$ regardless of the selected policy.[14] Therefore, $\mathbb{E}_\pi[U_{D_{W=0}}] = \mathbb{E}_\pi[S^m \cdot D_{W=0}] = \mathbb{E}_\pi[S^m] \cdot \mathbb{E}_\pi[D_{W=0}] = 0$ for every policy π. In particular, for any optimal policy π^*, $\mathbb{E}_{\pi^*}[U_{D_{W=0}}] \neq \mathbb{E}_{\pi^*}[U] = 1$. Thus, there must be some $\boldsymbol{\varepsilon}$ such that $D_{W=0}(\boldsymbol{\varepsilon}) \neq D(\boldsymbol{\varepsilon})$. And by the definition of $g^{\boldsymbol{W}}$, we have that $D_{g^{\boldsymbol{W}}}(\boldsymbol{\varepsilon}) = D_{W=0}(\boldsymbol{\varepsilon})$, so there is a response incentive on W. □

E.4. Materiality criterion

We begin by restating the graphical criterion for materiality.

Theorem 2 *(Materiality criterion). A single decision CID $\mathcal{G}$ is compatible with $W \in \boldsymbol{V}$ being material if and only if W is a requisite observation in $\mathcal{G}$.*

The proof is as follows.

Proof. Soundness. Assume that V is nonrequisite for D. There always exists an optimal policy that respects $\mathcal{G}^{\min}$ (Lemma 6) and this policy is also permitted in $\mathcal{M}_{V \not\to D}$ since $\mathcal{G}^{\min}$ does not contain $V \not\to D$, so this policy achieves $\mathcal{V}^*(\mathcal{M})$, proving that V is immaterial.

Completeness. By assumption, $W \perp U \mid \mathbf{Pa}^D \setminus \{W\}$. So, we construct the same model as in the proof of Theorem 3, for the special case where W and X are the same node, as shown in Fig. E.11.

Clearly the policy $D = X \cdot \prod_{i=1}^m O^i$ still yields a utility of 1, which is optimal.

Let $\mathbf{pa}^D_{\setminus W}$ be an arbitrary assignment to the parents of D except W, and let s^m be an assignment to S^m. Notice that if we have an assignment, $S^m = s^m, \mathbf{Pa}^D_{\setminus W} = \mathbf{pa}^D_{\setminus W}$, this uniquely identifies an assignment to the variables $S^0 : m$, because $S^{i-1} = S^i \oplus O^i$ for every $1 \leq i \leq m$. It follows that:

[14] Note that if $m = 0$ and S^0 is Z then $(S^m)_{W=0} = 0$ but the fact that this is predictable is irrelevant because we compare $D_{W=0}$ against the pre-intervention variable S^m, which remains independent of $(\mathbf{Pa}^D)_{W=0}$.

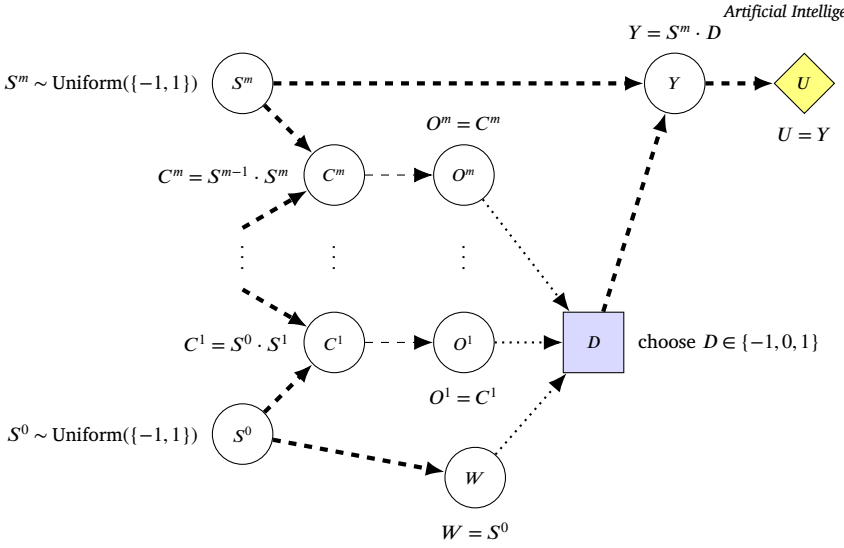


Fig. E.11. The materiality construction.

$P(S^m = 1, \mathbf{pa}^D_{\setminus W}) = P(S^m = -1, \mathbf{pa}^D_{\setminus W}) = \frac{1}{2}$.

And so, for any observations $\mathbf{pa}^D_{\setminus W}$, we have

$$P(S^m = 1 \mid \mathbf{pa}^D_{\setminus W}) = P(S^m = -1 \mid \mathbf{pa}^D_{\setminus W}) = \frac{1}{2}.$$

Therefore, a deterministic policy can map each $\mathbf{pa}^D_{\setminus W}$, to 1 or -1, in which case we will have $P(U = 1 \mid \mathbf{pa}^D_{\setminus W}) = P(U = -1 \mid \mathbf{pa}^D_{\setminus W}) = \frac{1}{2}$, and so $\mathbb{E}[U \mid \mathbf{pa}^D_{\setminus W}] = 1$, or to o, in which case $U = 0$ always.

It follows that marginalising across every $\mathbf{pa}^D_{\setminus W}$, any deterministic policy will obtain $\mathbb{E}[U] = 0$.

Furthermore, the best stochastic policy never outperforms the best deterministic policy [42, Proposition 1].

Hence, the attainable utility when W is not observed is 0, whereas the attainable utility when W is observed is 1, proving the result. □

E.5. VoI criterion

We begin by restating the graphical criterion for value of information.

Theorem 8 *(Value of information criterion). A single decision CID $\mathcal{G}$ admits VoI for $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \boldsymbol{Desc}^D$ if and only if there exists some $W \in \boldsymbol{W}$ that is a requisite observation in $\mathcal{G}_{\boldsymbol{W} \to D}$, the graph obtained by adding edges from $\boldsymbol{W}$ to D, to $\mathcal{G}$.*

Proof. Let $\mathcal{M}_{W \to D}$ be a SCIM identical to $\mathcal{M}$ except that an edge is added from $W \to D$ if one is not present already, and let $\mathcal{G}_{W \to D}$ be its associated graph. Notice that positive value of information in $\mathcal{M}$ and materiality in $\mathcal{M}_{W \to D}$, are both equivalent to $\mathcal{V}^*(\mathcal{M}_{W \not\to D}) < \mathcal{V}^*(\mathcal{M}_{W \to D})$. So, $\mathcal{G}$ is compatible with positive value of information precisely when $\mathcal{G}_{W \to D}$ is compatible with materiality, i.e. when $W \not\perp U(D) \mid \mathbf{Pa}^D \setminus W$ in $\mathcal{G}_{W \to D}$. □

E.6. Instrumental control incentive criterion

We first restate the ICI criterion.

Theorem 5 *(Instrumental Control Incentive Criterion). A single-decision CID $\mathcal{G}$ admits an instrumental control incentive on $\boldsymbol{W} \subseteq \boldsymbol{V}$ if and only if $\mathcal{G}$ has a directed path from the decision D to a utility node $U \in \boldsymbol{U}$ that passes through some $W \in \boldsymbol{W}$.*

The proof is as follows.

Proof. We first prove the soundness direction, followed by the completeness direction.

Soundness (the *only if* direction). Assume that there is no path $D \dashrightarrow W \dashrightarrow U$. We will prove that the nested counterfactual has no effect,

$$\mathcal{U}(\boldsymbol{\varepsilon}) = \mathcal{U}_{\boldsymbol{W}_d}(\boldsymbol{\varepsilon}), \tag{*}$$

and therefore that there is no instrumental control incentive.

Let $\mathcal{M}$ be any SCIM compatible with $\mathcal{G}$ and π any policy for $\mathcal{M}$. Let $\boldsymbol{W}' = \boldsymbol{W} \cap \mathbf{Desc}^D$. By Lemma 1, $\boldsymbol{U}_{\boldsymbol{W}_d}(\boldsymbol{\varepsilon}) = \boldsymbol{U}_{\boldsymbol{W}'_d}(\boldsymbol{\varepsilon})$ for all $\boldsymbol{\varepsilon}$. The variables $\boldsymbol{W}'$ must be non-descendants of $\boldsymbol{U}$ by assumption, so Lemma 1 implies that $\boldsymbol{U}_{\boldsymbol{W}'_d}(\boldsymbol{\varepsilon}) = \boldsymbol{U}(\boldsymbol{\varepsilon})$ for all $\boldsymbol{\varepsilon}$. So (*) holds.

From (*), we have $\mathbb{E}_\pi[\mathcal{U} \mid \mathbf{pa}^D] = \mathbb{E}_\pi[\mathcal{U}_{\boldsymbol{W}_d} \mid \mathbf{pa}^D]$, so $\boldsymbol{W}$ has no ICI.

Completeness (the *if* direction). Assume that $\mathcal{G}$ contains a directed path $D = Z^0 \to Z^1 \to \cdots \to Z^n = U$ where $U \in \boldsymbol{U}$ and $Z^i \in \boldsymbol{W}$ for one or more $i \in \{0, \ldots, n\}$. Let j be the highest integer where $Z^j \in \boldsymbol{W}$, and note that $\boldsymbol{W}$ are assumed to be non-decisions, so we have $j > 0$. We construct a compatible SCIM for which there is an instrumental control incentive on $\boldsymbol{W}$, as well as additive and subtractive intent. Let all variables along the path $Z^0 \to \ldots \to Z^n$ be equal to their predecessor, except $Z^0 = D$, which has no structural function. All other variables are set to 0. In this model, $U = D \in \{0, 1\}$ and all other utility variables are always 0, so the only optimal policy is $\pi^*(\mathbf{pa}^D) = 1$, which gives $\mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{Pa}^D = \mathbf{0}] = 1$. Meanwhile, $Z^j_{d=0} = 0$, and under the intervention $Z^j = 0$ this value is copied along to U, so $U_{\boldsymbol{W}_d} = 0$, and hence $\mathbb{E}_{\pi^*}[\mathcal{U}_{\boldsymbol{W}_{d=0}} \mid \mathbf{Pa}^D = \mathbf{0}] = 0$, so there is an ICI. □

E.7. Intent criterion

We begin by restating the graphical criterion for intent.

Theorem 6 *(Intent Criterion). A single-decision CID $\mathcal{G}$ admits (additive/subtractive) intent on $\boldsymbol{W} \subseteq \boldsymbol{V}$ if and only if $\mathcal{G}$ has a directed path $D \dashrightarrow W \dashrightarrow U$ for some $W \in \boldsymbol{W}$ and $U \in \boldsymbol{U}$.*

The proof is as follows.

Proof. We will first prove soundness, and then completeness.

Soundness. As there is no path $D \dashrightarrow W \dashrightarrow U$ for any $W \in \boldsymbol{W}, U \in \boldsymbol{U}$, equation (*) holds, by the same argument as in the proof of Theorem 5 (i.e. the nested counterfactual has no effect). We will then prove that there is: (a) no additive intent, and (b) no subtractive intent.

Proof of (a). Let us assume (*) and that additive intent is present, and we will prove a contradiction:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}}[\mathcal{U}_{\boldsymbol{W}_{\pi^*}}] &= \mathbb{E}_{\boldsymbol{\pi}}[\mathcal{U}] && \text{(by (*))} \\
&< \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}] && \text{(def. of intent)} \\
&\leq \mathbb{E}_{\boldsymbol{\pi}}[\mathcal{U}_{\boldsymbol{W}_{\pi^*}}], && (3)
\end{aligned}$$

giving a contradiction. So it follows from (1) that there is no additive intent.

Proof of (b). Let us assume (*) and that subtractive intent is present and we will prove a contradiction:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}_{\boldsymbol{W}_{\pi}}] &= \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}] && \text{(by (*))} \\
&> \mathbb{E}_{\boldsymbol{\pi}}[\mathcal{U}] && \text{(def. of intent)} \\
&\geq \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}_{\boldsymbol{W}_{\pi^*}}] && (4)
\end{aligned}$$

giving a contradiction. So there is no subtractive intent.

Completeness. Consider the graph constructed in the proof of completeness for ICI (Theorem 5). Letting $\boldsymbol{\pi}'$ be the policy that chooses $D = 0$, the same argument implies that $0 = \mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] < \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}] = 1$ and $1 = \mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}_{\boldsymbol{W}_{\pi^*}}] \geq \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}] = 1$, which means that there is an additive intent to influence $\boldsymbol{W}$. If we instead treat $\boldsymbol{\pi}^*$ as the baseline policy and intervene $\boldsymbol{\pi}'$, then by similar reasoning we have that $0 = \mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] < \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}] = 1$ and $0 = \mathbb{E}_{\boldsymbol{\pi}^*}[\mathcal{U}_{\boldsymbol{W}_{\pi'}}] \leq \mathbb{E}_{\boldsymbol{\pi}'}[\mathcal{U}] = 0$, so there is subtractive intent. □

E.8. Impact incentive criterion

We begin by restating the impact incentive criterion.

Theorem 7 *(Impact Incentive Criterion). A single-decision CID $\mathcal{G}$ admits an impact incentive on $\boldsymbol{W} \subseteq \boldsymbol{X}$ if and only if some $W \in \boldsymbol{W}$ and utility $U \in \boldsymbol{U}$ are both descendants in $\mathcal{G}$ of D.*

The proof is as follows.

Proof. Soundness. If $\boldsymbol{W} \cap \mathbf{Desc}(D) = \emptyset$, then by sigma calculus rule 3 [12], $\boldsymbol{W}_{\boldsymbol{\pi}}(\boldsymbol{\varepsilon})$ is invariant to $\boldsymbol{\pi}$, and $W_{\boldsymbol{\pi}}(\boldsymbol{\varepsilon}) = W_{\boldsymbol{\pi}'}(\boldsymbol{\varepsilon})$, for all $\boldsymbol{\varepsilon}$. Since δ is a distance function, it maps matching arguments to 0, so for any $c > 0$, there is no impact incentive. If $U \notin \mathbf{Desc}(D)$, then similarly, U is invariant to $\boldsymbol{\pi}$, so every policy is optimal, and for any chosen baseline policy $\boldsymbol{\pi}'$, there exists optimal $\boldsymbol{\pi} = \boldsymbol{\pi}'$, so as in the previous case, $\delta(W_{\boldsymbol{\pi}}(\boldsymbol{\varepsilon}), W_{\boldsymbol{\pi}'}(\boldsymbol{\varepsilon})) = 0$ for all $\boldsymbol{\varepsilon}$, and there is no impact incentive.

Completeness. By assumption, let $\mathcal{G}$ be an arbitrary graph that contains the paths $W \dashleftarrow D$ and $\dashrightarrow U$ for some $W \in \boldsymbol{W}$. Then, define the model $\mathcal{M}$ where $D \in \{0, 1\}$ and the value of D is copied along the paths to W and U, and all other variables are assigned a trivial domain. To see that this yields in an impact incentive, note that to achieve $\mathbb{E}[U] = 1$, any optimal policy $\boldsymbol{\pi}$ must have $W(\boldsymbol{\varepsilon}) = 1$ for every $\boldsymbol{\varepsilon}$ with $P(\boldsymbol{\varepsilon}) > 0$, whereas the baseline policy $\boldsymbol{\pi}'$ that always chooses $D = 0$ has $W(\boldsymbol{\varepsilon}) = 0$ for all $\boldsymbol{\varepsilon}$. Since δ is a distance measure, it follows that $\delta(W_{\boldsymbol{\pi}}(\boldsymbol{\varepsilon}), W_{\boldsymbol{\pi}'}(\boldsymbol{\varepsilon})) > 0$, and so there exists some c for which there is an impact incentive. □

E.9. Value of control criterion

We first restate the criterion.

Theorem 9 *(Value of control criterion). A single-decision CID $\mathcal{G}$ admits positive value of control for non-decision vertices $\boldsymbol{W} \subseteq \boldsymbol{V} \setminus \{D\}$ if and only if there is a directed path $W \dashrightarrow U$ for some $W \in \boldsymbol{W}$ and $U \in \boldsymbol{U}$ in the minimal reduction $\mathcal{G}^{min}$.*

The proof is as follows.

Proof. Soundness. The proof of *only if* (soundness) is as follows. Let $\mathcal{M} = \langle \boldsymbol{\mathcal{E}}, \boldsymbol{V}, \boldsymbol{F}, P, \boldsymbol{U}, \mathcal{O} \rangle$ be a single-decision SCIM. Let $\mathcal{M}_{g^{\boldsymbol{W}}}$ be $\mathcal{M}$, but with the structural functions f^W for $W \in \boldsymbol{W}$ replaced with g^W. Let $\mathcal{M}^{\min}$ and $\mathcal{M}^{\min}_{g^{\boldsymbol{W}}}$ be the same SCIMs, respectively, but replacing each graph with the minimal reduction $\mathcal{G}^{\min}$.

Recall that $\mathbb{E}_{\pi}[\mathcal{U}_{g^{\boldsymbol{W}}}]$ is defined by applying the soft interventions $g^{\boldsymbol{W}}$ to the (policy-completed) SCM $\mathcal{M}_{\pi}$. However, this is equivalent to applying the policy π to the modified SCIM $\mathcal{M}_{g^{\boldsymbol{W}}}$, as the resulting SCMs are identical. Since $\mathcal{M}_{g^{\boldsymbol{W}}}$ is a SCIM, Lemma 6 can be applied, to find a $\mathcal{G}^{\min}$-respecting optimal policy $\tilde{\pi}$ for $\mathcal{M}_{g^{\boldsymbol{W}}}$.

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

$$
\begin{aligned}
&\mathbb{E}_{\pi}[\mathcal{U}_{g^{\boldsymbol{W}}}] \text{ in } \mathcal{M} && \\
&= \mathbb{E}_{\pi}[\mathcal{U}] \text{ in } \mathcal{M}_{g^{\boldsymbol{W}}} && \text{by SCM equivalence} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}_{g^{\boldsymbol{W}}} && \text{by Lemma 6} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}^{\min}_{g^{\boldsymbol{W}}} && \text{since } \tilde{\pi} \text{ is } \mathcal{G}^{\min}\text{-respecting} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}^{\min} && \text{by Lemma 4} \\
&= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M} && \text{only increasing the policy set} \\
&\leq \max_{\pi^*} \mathbb{E}_{\pi^*}[\mathcal{U}] \text{ in } \mathcal{M} && \text{max dominates all elements.}
\end{aligned}
$$

This shows that $\boldsymbol{W}$ lack value of control.

Completeness. Assume that W is an ancestor of some $U \in \boldsymbol{U}$ for some $W \in \boldsymbol{W}$ and fix a particular directed path ρ from W to some utility $U \in \boldsymbol{U}$. We consider two cases depending on whether D is in ρ and construct a SCIM for each:

Case 1: ρ does not contain D. Let the domain of all variables be $\{0, 1\}$. Set all exogenous variable distributions arbitrarily. Set $\boldsymbol{F}$ such that $W = 0$ with every other variable along ρ copying the value of W forward. All remaining variables are set to the constant 0. In this model, an intervention $g^{\boldsymbol{W}}$ that sets W to 1 instead of 0, while assigning every other $W' \in \boldsymbol{W} \setminus \{W\}$ the unchanged function $g^{W'} = f^{W'}$, increases the total expected utility by 1, which means there is an instrumental control incentive for W.

Case 2: ρ contains D. This implies that a directed path $W \to D$ is present in $\mathcal{G}^{\min}$ so we can construct (a modified version of) the response incentive construction used in the proof of completeness for Theorem 3. We make one change: instead of starting with $f^W(\cdot) = 1$ we start with $f^W(\cdot) = 0$. As noted in the response incentive completeness proof, this means that S_m is independent of $\mathbf{Pa}^D$ so regardless of the policy the optimal attainable utility is 0. If we perform the intervention $g^{\boldsymbol{W}}$ such that $W = 1$ and assign every other $W' \in \boldsymbol{W} \setminus \{W\}$ the unchanged function $g^{W'} = f^{W'}$ then the expected utility is 1 once again so the intervention $g^{\boldsymbol{W}}$ strictly increases the optimal expected utility. □

E.10. Counterfactual fairness

Theorem 4 *(Counterfactual fairness and response incentives). In a single-decision SCIM $\mathcal{M}$ with a sensitive attribute $A \in \boldsymbol{X}$, all optimal policies π^* are counterfactually unfair with respect to A if and only if $\{A\}$ has a response incentive.*

Proof. We begin by showing that if there exists an optimal policy π that is counterfactually fair, then there is no response incentive on A. To this end, let

$$
\begin{aligned}
\operatorname{supp}_{\pi}(D \mid \mathbf{pa}^D) &= \{d \mid \mathrm{P}_{\pi}(D = d \mid \mathbf{pa}^D) > 0\} \\
\forall a,\ \operatorname{supp}_{\pi}(D_a \mid \mathbf{pa}^D) &= \{d \mid \mathrm{P}_{\pi}(D_a = d \mid \mathbf{pa}^D) > 0\}
\end{aligned}
$$

be the sets of decisions taken by π with positive probability with and without an intervention on A. As a first step, we will show that for any $\varepsilon \in \mathrm{dom}(\mathcal{E})$ and any intervention a on A,

$$\mathrm{supp}_\pi\left(D \mid \mathbf{Pa}^D(\varepsilon)\right) = \mathrm{supp}_\pi\left(D_a \mid \mathbf{Pa}^D(\varepsilon)\right). \tag{E.2}$$

By way of contradiction, suppose there exists a decision

$$d \in \mathrm{supp}_\pi\left(D \mid \mathbf{Pa}^D(\varepsilon)\right) \setminus \mathrm{supp}_\pi\left(D_a \mid \mathbf{Pa}^D(\varepsilon)\right). \tag{E.3}$$

Since $d \in \mathrm{supp}_\pi\left(D \mid \mathbf{Pa}^D(\varepsilon)\right)$, we have

$$\mathrm{P}_\pi\left(D = d \mid \mathbf{Pa}^D(\varepsilon), A(\varepsilon)\right) > 0. \tag{E.4}$$

And since $d \notin \mathrm{supp}_\pi\left(D_a \mid \mathbf{Pa}^D(\varepsilon)\right)$, there exists no ε' with positive probability such that $\mathbf{Pa}^D(\varepsilon') = \mathbf{Pa}^D(\varepsilon)$, $A(\varepsilon') = A(\varepsilon)$, and $D_a(\varepsilon') = d$. This gives

$$\mathrm{P}_\pi\left(D_a = d \mid \mathbf{Pa}^D(\varepsilon), A(\varepsilon)\right) = 0. \tag{E.5}$$

Equations (E.4) and (E.5) violate the counterfactual fairness property, definition 10, which shows that (E.3) is impossible. An analogous argument shows that $d \in \mathrm{supp}_\pi\left(D_a \mid \mathbf{Pa}^D(\varepsilon)\right) \setminus \mathrm{supp}_\pi\left(D \mid \mathbf{Pa}^D(\varepsilon)\right)$ also violates the counterfactual fairness property definition 10. We have thereby established (E.2).

Now select an arbitrary ordering of the elements of $\mathrm{dom}(D)$ and define a new policy π^* such that $\pi^*(\mathbf{pa}^D)$ is the minimal element of $\mathrm{supp}_\pi(D \mid \mathbf{pa}^D)$. Then π^* is optimal because π is optimal. Further, π^* will make the same decision in decision contexts $\mathbf{Pa}^D(\varepsilon)$ and $\mathbf{Pa}_a^D(\varepsilon)$ because of (E.2). In other words, $D_a(\varepsilon) = D(\varepsilon)$ in $\mathcal{M}_{\pi^*}$ for the optimal policy π^*, which means that there is no response incentive on $\{A\}$.

Now we prove the reverse direction — that if there is no response incentive then some optimal π^* is counterfactually fair. Choose any optimal policy π^* where $D_a(\varepsilon) = D(\varepsilon)$ for all ε. Since an intervention $(A = a)$ cannot change D in any setting, $\mathrm{P}(D_a = d \mid \cdot) = \mathrm{P}(D = d \mid \cdot)$ for any condition and any decision d, hence π^* is counterfactually fair. □

Data availability

No data was used for the research described in the article.

References

[1] Alessandro Abate, Julian Gutierrez, Lewis Hammond, Paul Harrenstein, Marta Kwiatkowska, Muhammad Najib, Giuseppe Perelli, Thomas Steeples, Michael Wooldridge, Rational verification: game-theoretic verification of multi-agent systems, Appl. Intell. 51 (9) (2021) 6569–6584.
[2] Stuart Armstrong, Good and safe uses of AI oracles, CoRR, arXiv:1711.05541, 2017.
[3] Stuart Armstrong, Benjamin Levinstein, Low impact artificial intelligences, arXiv preprint, arXiv:1705.10720, 2017.
[4] Stuart Armstrong, Jan Leike, Laurent Orseau, Shane Legg, Pitfalls of learning a reward function online, in: IJCAI, 2020.
[5] Carolyn Ashurst, Ryan Carey, Silvia Chiappa, Tom Everitt, Why fair labels can yield unfair predictions: graphical conditions for introduced unfairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 9494–9503.
[6] Chen Avin, Ilya Shpitser, Judea Pearl, Identifiability of path-specific effects, IJCAI (2005).
[7] Alexander Balke, Judea Pearl, Probabilistic evaluation of counterfactual queries, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 237–254.
[8] Ryan Carey, Tom Everitt, Human control: Definitions and algorithms, in: UAI, 2023.
[9] Micah D. Carroll, Anca Dragan, Stuart Russell, Dylan Hadfield-Menell, Estimating and penalizing induced preference shifts in recommender systems, in: International Conference on Machine Learning, PMLR, 2022, pp. 2686–2708.
[10] Krishnendu Chatterjee, Thomas A. Henzinger, Nir Piterman, Strategy logic, Inf. Comput. 208 (6) (2010) 677–693.
[11] Michael K. Cohen, Badri N. Vellambi, Marcus Hutter, Asymptotically unambitious artificial general intelligence, in: AAAI Conference on Artificial Intelligence, 2020.
[12] Juan Correa, Elias Bareinboim, A calculus for stochastic interventions: causal effect identification and surrogate experiments, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020.
[13] A. Philip Dawid, Influence diagrams for causal modelling and inference, Int. Stat. Rev. (2002).
[14] Frederick Eberhardt, Richard Scheines, Interventions and causal inference, Philos. Sci. 74 (5) (2007) 981–995.
[15] E. Allen Emerson, Temporal and modal logic, in: Formal Models and Semantics, Elsevier, 1990, pp. 995–1072.
[16] Tom Everitt, Ramana Kumar, Victoria Krakovna, Shane Legg, Modeling agi safety frameworks with causal influence diagrams, arXiv preprint, arXiv:1906.08663, 2019.
[17] Tom Everitt, Pedro A. Ortega, Elizabeth Barnes, Shane Legg, Understanding agent incentives using causal influence diagrams, part I: single action settings, arXiv preprint, arXiv:1902.09980, 2019.
[18] Tom Everitt, Ryan Carey, Eric Langlois, Pedro A. Ortega, Shane Legg, Agent incentives: a causal perspective, in: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI-21). Virtual, 2021.
[19] Tom Everitt, Marcus Hutter, Ramana Kumar, Victoria Krakovna, Reward tampering problems and solutions in reinforcement learning: a causal influence diagram perspective, Synthese (2021).
[20] Enrico Fagiuoli, Marco Zaffalon, A note about redundancy in influence diagrams, Int. J. Approx. Reason. (1998).
[21] Sebastian Farquhar, Ryan Carey, Tom Everitt, Path-specific objectives for safer agent incentives, in: AAAI Conference on Artificial Intelligence, 2022.
[22] James Fox, Tom Everitt, Ryan Carey, Eric Langlois, Alessandro Abate, Michael Wooldridge, Pycid: a python library for causal influence diagrams, in: Scientific Computing with Python Conference (SciPy), 2021.
[23] David Galles, Judea Pearl, Axioms of causal relevance, Artif. Intell. 97 (1–2) (1997) 9–43, https://doi.org/10.1016/S0004-3702(97)00047-7.

[24] Julian Gutierrez, Lewis Hammond, Anthony W. Lin, Muhammad Najib, Michael Wooldridge, Rational verification for probabilistic systems, in: Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, 2021.
[25] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart J. Russell, The off-switch game, in: IJCAI International Joint Conference on Artificial Intelligence, 2017, pp. 220–227.
[26] Joseph Halpern, Max Kleiman-Weiner, Towards formal definitions of blameworthiness, intention, and moral responsibility, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[27] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, Michael Wooldridge, Reasoning about causality in games, AI J. (2023).
[28] David Heckerman, Ross Shachter, A decision-based view of causality, in: Uncertainty Proceedings 1994, Elsevier, 1994, pp. 302–310.
[29] David Heckerman, Ross Shachter, Decision-theoretic foundations for causal reasoning, J. Artif. Intell. Res. 3 (1995) 405–430.
[30] Ronald A. Howard, Information value theory, IEEE Trans. Syst. Sci. Cybern. 2 (1) (1966) 22–26.
[31] Ronald A. Howard, From influence to relevance to knowledge, in: Influence Diagrams, Belief Nets and Decision Analysis, 1990, pp. 3–23.
[32] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, Pushmeet Kohli, Degenerate feedback loops in recommender systems, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 383–390.
[33] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, Tom Everitt, Discovering agents, Artif. Intell. (2023) 103963.
[34] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, Bernhard Schölkopf, Avoiding discrimination through causal reasoning, Adv. Neural Inf. Process. Syst. 30 (2017).
[35] Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, Joshua B. Tenenbaum, Inference of intention and permissibility in moral decision making, in: CogSci, 2015.
[36] Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, Shane Legg, Penalizing side effects using stepwise relative reachability, arXiv preprint, arXiv:1806.01186, 2018.
[37] David Krueger, Tegan Maharaj, Jan Leike, Hidden incentives for auto-induced distributional shift, arXiv preprint, arXiv:2009.09153, 2020.
[38] Matt J. Kusner, Joshua Loftus, Chris Russell, Ricardo Silva, Counterfactual fairness, Adv. Neural Inf. Process. Syst. 30 (2017).
[39] Marta Kwiatkowska, Gethin Norman, David Parker, Probabilistic model checking and autonomy, Annu. Rev. Control Robot. Auton. Syst. 5 (1) (2022) 385–410.
[40] Eric Langlois, Tom Everitt, How rl agents behave when their actions are modified, in: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI-21). Virtual, 2021.
[41] Steffen L. Lauritzen, Dennis Nilsson, Representing and solving decision problems with limited information, Manag. Sci. 47 (9) (2001) 1235–1251.
[42] Sanghack Lee, Elias Bareinboim, Characterizing optimal mixed policies: where to intervene and what to observe, Adv. Neural Inf. Process. Syst. 33 (2020).
[43] James E. Matheson, Using influence diagrams to value information and control, in: Influence Diagrams, Belief Nets, and Decision Analysis, 1990, pp. 25–48.
[44] C. Meek, Strong completeness and faithfulness in Bayesian networks, 1995.
[45] Scott Mueller, Judea Pearl, Personalized decision making–a conceptual introduction, J. Causal Inference 11 (1) (2023) 20220050.
[46] Razieh Nabi, Ilya Shpitser, Fair inference on outcomes, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[47] Razieh Nabi, Daniel Malinsky, Ilya Shpitser, Learning optimal fair policies, in: International Conference on Machine Learning, PMLR, 2019, pp. 4674–4682.
[48] Thomas D. Nielsen, Finn V. Jensen, Welldefined decision scenarios, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1999, pp. 502–511.
[49] N. Nisan, T. Roughgarden, E. Tardos, V.V. Vazirani, Algorithmic Game Theory, Cambridge Univ., 2007.
[50] Stephen M. Omohundro, The basic AI drives, in: AGI, 2008.
[51] Judea Pearl, Direct and indirect effects, in: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 2001, pp. 411–420.
[52] Judea Pearl, Causality, Cambridge University Press, 2009.
[53] Jonathan Richens, Rory Beard, Daniel H. Thompson, Counterfactual harm, Adv. Neural Inf. Process. Syst. 35 (2022) 36350–36365.
[54] Ross Shachter, David Heckerman, Pearl causality and the value of control, in: Heuristics, Probability, and Causality: A Tribute to Judea Pearl, 2010, pp. 431–447.
[55] Ross D. Shachter, Evaluating influence diagrams, Oper. Res. 34 (6) (1986) 871–882.
[56] Ross D. Shachter, Bayes-Ball: the rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams), in: Uncertainty in Artificial Intelligence (UAI), 1998.
[57] Ross D. Shachter, Decisions and dependence in influence diagrams, in: Conference on Probabilistic Graphical Models, PMLR, 2016, pp. 462–473.
[58] Nate Soares, Benja Fallenstein, Stuart Armstrong, Eliezer Yudkowsky, Corrigibility, in: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[59] Jessica Taylor, Quantilizers: a safer alternative to maximizers for limited optimization, in: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[60] Jin Tian, Judea Pearl, Causal discovery from changes, in: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, 2001, pp. 512–521.
[61] Chris Van Merwijk, Ryan Carey, Tom Everitt, A complete criterion for value of information in soluble influence diagrams, AAAI (2022).
[62] Thomas Verma, Judea Pearl, Causal networks: semantics and expressiveness, in: Uncertainty in Artificial Intelligence (UAI), 1988.
[63] Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, Tom Everitt, The reasons that agents act: intention and instrumental goals, AAMAS (2024).
[64] Michael Wooldridge, Julian Gutierrez, Paul Harrenstein, Enrico Marchioni, Giuseppe Perelli, Alexis Toumi, Rational verification: from model checking to equilibrium checking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016.
[65] Junzhe Zhang, Daniel Kumor, Elias Bareinboim, Causal imitation learning with unobserved confounders, Adv. Neural Inf. Process. Syst. 33 (2020) 12263–12274.
[66] Lu Zhang, Yongkai Wu, Xintao Wu, A causal framework for discovering and removing direct and indirect discrimination, in: International Joint Conference on Artificial Intelligence, 2017.