

Incentives and AI Safety in Causal Models

Ryan Carey

Jesus College

University of Oxford

*A thesis submitted for the degree of
Doctor of Philosophy in Statistics*

June 2024

Abstract

One kind of concern about AI is that sufficiently advanced AI systems might outstrip the control of their programmers and users. At the core of this concern is that it is often hard to specify a notion of success for autonomous decision-making tasks that is perfectly accurate, and instead, an assigned objective function will often incentivise some kind of undesired behaviour. In the worst case, one might worry that an AI system might be incentivised to ignore or manipulate shutdown instructions, in order to keep pursuing a misspecified goal. One framework well-suited to modelling agents' incentives is the causal influence diagram, a graphical model that contains decision and utility nodes, and edges that denote causal relationships. This thesis will explore, in three main parts, how causal notions of incentives can be applied to AI safety, and evaluated based on the graphical structure. The first part defines some classes of incentives: response incentives indicate how an optimal agent causally responds to aspects of its environment (such as commands), instrumental control incentives indicate an agent's tendency to gain utility by influencing parts of the environment, and impact incentives indicate the tendency to influence the environment, whether this is instrumental to utility-attainment, or a mere side-effect. It is often possible to rule out a possible incentive based on the graphical structure alone. The thesis will present graphical criteria for ruling out incentives in single decision settings, which are complete (i.e. they rule out the incentives using the graphical structure whenever possible). The second part formalises the problem of safe agent shutdown, using a class of causal influence diagrams called shutdown games. In a shutdown game, an AI system is guaranteed not to harm a human overseer, if it is obedient and cautious, while the overseer is vigilant. The third part is a step toward a complete criterion for materiality. An observation is said to be material if making it unobservable would reduce the attainable expected utility. I show that in many graphs, where the strongest existing criterion cannot prove immateriality, materiality is indeed possible, bringing us one step closer to a complete criterion. Finally, I discuss how causal models of incentives succeed and fail as a model of AI safety.

Incentives and AI Safety in Causal Models



Ryan Carey
Jesus College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Statistics

June 2024

Acknowledgements

A heartfelt thank you to my supervisor Robin for the many research discussions that have deepened my appreciation of causal models. I have been incredibly lucky to have a supervisor who allowed me the freedom to work on all kinds of projects, while offering invaluable comments and improvements on all of them.

To Tom Everitt: I remember clearly our discussions in 2019 at FHI about Pearl's causal models, and how it seemed like they might speak to problems in AI safety. I'm very proud of how far this research direction has come since that time, and glad to have shared this journey with you.

To Nick Bostrom, thank you for offering me an intellectual sanctuary at Oxford before I began my DPhil studies. To my collaborators in the research direction that this thesis concentrates on — Sanghack Lee, Chris van Merwijk, Eric Langlois, James Fox, Lewis Hammond, Sebastian Farquhar, Carolyn Ashurst — it has been a joy to work with you. Thank you for enriching each of our projects. To Open Philanthropy, thank you for the generous funding that made my studies possible.

To all of my family and loved ones, your unwavering support has enabled me to pursue this dream. A special thank you to my parents Peter and Brenda, for their unconditional love and support, for nurturing intellectual curiosity throughout my education, and for teaching me so much of what I know.

Without assistance and encouragement from all of you, this thesis would not have been possible.

Abstract

One kind of concern about AI is that sufficiently advanced AI systems might outstrip the control of their programmers and users. At the core of this concern is that it is often hard to specify a notion of success for autonomous decision-making tasks that is perfectly accurate, and instead, an assigned objective function will often incentivise some kind of undesired behaviour. In the worst case, one might worry that an AI system might be incentivised to ignore or manipulate shutdown instructions, in order to keep pursuing a misspecified goal. One framework well-suited to modelling agents' incentives is the causal influence diagram, a graphical model that contains decision and utility nodes, and edges that denote causal relationships. This thesis will explore, in three main parts, how causal notions of incentives can be applied to AI safety, and evaluated based on the graphical structure. The first part defines some classes of incentives: response incentives indicate how an optimal agent causally responds to aspects of its environment (such as commands), instrumental control incentives indicate an agent's tendency to gain utility by influencing parts of the environment, and impact incentives indicate the tendency to influence the environment, whether this is instrumental to utility-attainment, or a mere side-effect. It is often possible to rule out a possible incentive based on the graphical structure alone. The thesis will present graphical criteria for ruling out incentives in single decision settings, which are complete (i.e. they rule out the incentives using the graphical structure whenever possible). The second part formalises the problem of safe agent shutdown, using a class of causal influence diagrams called shutdown games. In a shutdown game, an AI system is guaranteed not to harm a human overseer, if it is obedient and cautious, while the overseer is vigilant. The third part is a step toward a complete criterion for materiality. An observation is said to be material if making it unobservable would reduce the attainable expected utility. I show that in many graphs, where the strongest existing criterion cannot prove immateriality, materiality is indeed possible, bringing us one step closer to a complete criterion. Finally, I discuss how causal models of incentives succeed and fail as a model of AI safety.

Contents

List of Figures	vi
1 Introduction	2
1.1 Why address AI safety and control using causal models of incentives?	2
1.2 The scope of this thesis	5
1.3 Contributions and organisation of this thesis	6
2 Background	9
2.1 Causal models of agents	9
2.2 AI safety and fairness	18
2.3 Graphical models of incentives	22
3 Incentives for Responsiveness, Instrumental Control and Impact	28
3.1 Introduction	29
3.2 Setup	33
3.3 Materiality	39
3.4 Response incentives	41
3.5 Instrumental control incentives	44
3.6 Intent	48
3.7 Impact incentives	51
3.8 Incentives in a multi-decision detting	55
3.9 Related work	57
3.10 Discussion and conclusion	60
4 Human Control: Definitions and Algorithms	65
4.1 Introduction	66
4.2 Literature review	67
4.3 Structural causal influence models	69
4.4 Shutdown problem	71
4.5 Routes to control	73
4.6 Algorithms	82
4.7 Discussion	86

5	Toward a Complete Criterion for Value of Information in Insoluble Decision Problems	91
5.1	Introduction	92
5.2	Setup	96
5.3	Review of graphical criteria for materiality	100
5.4	Main result	106
5.5	Toward a more general proof of materiality	125
5.6	Conclusion	129
5.7	Acknowledgements	130
6	Discussion and Conclusions	132
6.1	The contributions of this thesis	132
6.2	Limitations	135
6.3	Conclusion	138
A	Incentives for Responsiveness, Instrumental Control and Impact (Supplementary Materials)	139
B	Human Control: Definitions and Algorithms (Supplementary Materials)	157
C	Toward a Complete Criterion for Value of Information in Insoluble Decision Problems (Supplementary Materials)	170
	References	187

List of Figures

2.1	A Bayesian network representing a sprinkler and grass growth . . .	12
2.2	A pair of SCMs with the same interventional distributions	13
2.3	Venn diagram of modelling frameworks	15
2.4	An influence diagram and a structural causal influence diagram . .	15
2.5	Soluble and insoluble influence diagrams	25
3.1	Grade predictor and content recommendation examples	32
3.2	Interventions in a structural causal influence model	35
3.3	Incentives for the grade-predictor	43
3.4	Content recommendation incentives	46
3.5	An impact measure, illustrated in a twin graph	53
3.6	Incentives of a Q-learner	55
3.7	The task of opening a combination lock	56
4.1	Shutdown problem example	67
4.2	General shutdown problem	72
4.3	Shutdown problem with indifference methods	85
5.1	Materiality examples	94
5.2	Immaterial decision example	100
5.3	Past materiality constructions	102
5.4	Materiality in soluble graphs	102
5.5	Two SCMs, with models constructed using different (red) info paths.	108
5.6	The materiality paths	111
5.7	Constructing the model for the info path	114
5.8	The materiality SCM	117
5.9	The intersection node is, or is not, a decision in the proof of materiality	119
5.10	Randomness may be needed to prove materiality	128
5.11	Another materiality counterexample: superimposition	129

Notation

Symbol	Object	Page
$\mathbf{Anc}(X)$,	Ancestors	96
$\llbracket P \rrbracket$	Iverson brackets	84
$[\mathbf{Z}]$	Implied variables	100
$\langle x, y \rangle$	Tuple	97
\mathbf{C}^X	Contexts	98
$\mathbf{Ch}(X)$	Children	96
\mathbf{Desc}^V or $\mathbf{Desc}(X)$	Descendants of V	36, 96
do	Do operator	10, 34
$\text{dom}(V)$ or \mathfrak{X}_V	Domain of V	37
\mathcal{E}_V	Exogenous variable for V	12,37
$\text{exp}_2^n(k)$	Repeated exponent	114
\mathbf{Fa}^V	Family of V	70
X_{Y_z}	Nested potential response	35
$\mathcal{G}(\mathcal{S})$	Scoped graph	98
g^X	Soft intervention	34
\mathcal{G}^{\min}	Minimal reduction	41
\mathcal{M}_x	Sub-model	34
$p_{1:N}$	a sequence of paths p_1, \dots, p_N	97
\mathbf{Pa}^V or $\mathbf{Pa}(X)$	Parents of V	36, 96
$\mathbf{Pa}(V^p)$	p -component of parent of V along p	116
\perp or $\perp_{\mathcal{G}}$	d-separation	36,99
$<$	An ordering over variables	105
P^π	Probability under policy π	38
π	A policy	38
\mathcal{S}	Mixed policy scope	97
\mathcal{U}_x	Total utility under intervention	38
$U(\mathbb{B}^n)$	Uniform distribution over $\{0, 1\}^n$	97
V	Variable	34
V^p	Component of V for path p	116
\mathbf{V}	Set of variables	34
v	Assignment to a variable	34
\mathbf{v}	Assignment to a set of variables	34
$\mathbf{V}_{1:N}$	a sequence of variables	97
$\mathbf{v}_{1:N}$	a sequence of assignments	97
$\mathcal{V}^*(\mathcal{M})$	Maximum expected utility	39
$\mathbf{X}(\mathcal{S})$	Decisions in a scope	97
$x[y]$	Indexing	97
\oplus	XOR operator	13

1

Introduction

Contents

1.1	Why address AI safety and control using causal models of incentives?	2
1.2	The scope of this thesis	5
1.3	Contributions and organisation of this thesis	6

1.1 Why address AI safety and control using causal models of incentives?

The development of increasingly capable artificial intelligence presents numerous potential risks and benefits. AI could substantially accelerate economic growth [Chui et al., 2023, Trammell and Korinek, 2023] and innovation [Cockburn et al., 2018] yet it also raises concerns about trustworthiness [Mirsky and Lee, 2021], interpretability [Linardatos et al., 2020], fairness [Barocas et al., 2023], privacy [Oseni et al., 2021], technological unemployment [Webb, 2019], and lethal autonomous weapons [Horowitz, 2021]. Some have argued that eventually, humanity may even struggle to maintain control of increasingly capable AI systems [Bostrom, 2014, Bengio et al., 2023, Dalrymple et al., 2024].

Safety researchers have begun to devise specifications for what fair and trustworthy AI behaviour should look like. In many cases, notions of causality play a role in defining these specifications. In fairness, for example, some of the earliest specifications, such as demographic parity, were limited to the statistical properties of AI or human decisions [Barocas et al., 2023]. Over time, however, it has been proposed that systems ought to be counterfactually fair [Kusner et al., 2017b], in that sensitive demographic variables ought not causally affect the decision, or at least they ought not do so along certain causal pathways [Nabi and Shpitser, 2018, Chiappa, 2019]. Once causal fairness specifications are devised, algorithms can be designed to respect them [Nabi and Shpitser, 2018, Nabi et al., 2022]. On matters of interpretability too, over time, there has been an increased interest in deriving “mechanistic” (or causal) models of the internals of AI systems, in order to understand their behaviour [Bereska and Gavves, 2024].

Over time, AI systems have become more capable, which may allow them to pose greater risks. Recently, large language models have demonstrated near-human-level performance in coding, medical, and mathematical tasks, and many other areas [Bubeck et al., 2023]. They have demonstrated some risky capabilities, such as an ability to search a file system for a password, or place a phone call, but lacked others, such as the ability to create a bitcoin wallet, or create a new language model agent [Kinniment et al., 2023]. Extrapolating current trends, many other risky capabilities may be unlocked in the future. Indeed, experts now offer a median prediction that AI systems will outperform humans at all work tasks within three decades, an estimate that is 13 years earlier than in a similar survey performed one year prior [Grace et al., 2024]. For open-ended decision-making tasks, it is often difficult to define a goal that describes success at that task with perfect accuracy [Dalrymple et al., 2024, Krakovna et al., 2020b]. So, one concern is that when highly-capable AI systems are assigned open-ended tasks, using imperfect utility functions, this may incentivise behaviour that is contrary to implicit goals of designers. In the most extreme scenario, such a system could have an incentive to disregard a shutdown

instruction, or prevent it from being given [Omohundro, 2008, Soares et al., 2015].

There is substantial disagreement about whether control is a serious and realistic issue for AI systems. If there is a real risk that humanity may fail to control AI systems, then this would seem extremely important, and even a potential existential threat to humanity [Bostrom, 2003, 2013]. In line with this, prominent CEOs and professors recently signed a letter asserting that “[m]itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.” [Center for AI Safety, 2023] Some experts, however, view these risks as unrealistic or as a distraction from other concerns like ensuring that AI systems behave in a fair and interpretable manner [Crawford and Calo, 2016, Etzioni, 2016]. Consensus on this question is difficult to achieve, because the case for risk relies on speculation about the capabilities of future systems.

An important question is whether we can specify how controllable AI ought to behave, similar to previous specifications for AI fairness. Specifications for controllability could be used in the design of AI systems. They also could be used to discern whether controllable AI is possible or likely, based on analysing whether these specifications could be met in plausible environments. Such specifications could also be compared with fairness specifications, in order to arrive at general principles for safe AI design.

Notions of controllability, like fairness, often have a causal character: When we say that a system should be controllable by a human overseer, this presumably means that the overseer’s instructions exert some causal effect on the agent’s actions. There also exist kinds of causal models that can describe an agent’s incentives. Causal influence diagrams highlight certain variables as decisions, and others as utility variables, that the agent’s policy is chosen to achieve [Dawid, 2002, Lee and Bareinboim, 2020]. These models are also endowed with notions of incentives. Certain observations are deemed *material*, meaning that their value must be taken into account, in order to make an optimal decision. Causal models can also be used to describe which variables an agent *intends* to influence, based on whether the

influence of those outcomes is the reason that a policy is optimal [Kleiman-Weiner et al., 2015, Halpern and Kleiman-Weiner, 2018].¹.

So, the overarching goal of this thesis will be to use causal models of agents to define incentive concepts that describe safe and unsafe AI behaviour, and especially the controllability of AI systems. In detail, the hope is to:

- introduce tools that enable matters of AI safety to be understood in a more precise, organised, and clear way, such as causal modelling frameworks, notions of incentives, and their graphical criteria (which indicate when these incentives may arise);
- demonstrate avenues for further theoretical work on AI safety, using incentive concepts; and
- explore how incentive concepts relate to fairness, and safety, so that they may be used as specifications for future agent designs, or legal or ethical analyses.

1.2 The scope of this thesis

It is useful to clarify the distinction between *causal models of agents*, and *causal models of incentives*. In my usage, the first will refer to causal models whose variables are adapted for their use in modelling agents, by partitioning the variables into decisions, utilities and regular variables. The second refers to concepts such as intent and incentives, that may be defined in these models based on the agent’s utility and causal relationships. The focus of this thesis is causal models of incentives, and their application to AI safety. The primary application of interest is AI control. Other safety and fairness issues are also discussed, but other kinds of risks from AI systems (like technological unemployment) are outside of the scope of this thesis.

The approach used in this investigation is conceptual and theoretical, and so although it includes formal worked examples that explain possible AI applications, it does not include any actual AI experiments.

¹Different formalisations of intent are further discussed in Chapter 3.

1.3 Contributions and organisation of this thesis

Apart from this introduction, the thesis comprises five chapters and includes three research papers. I now outline their content, and how they will address the overarching goals of the thesis.

Background In Chapter 2, I will review causal models and how they are used to model agents. I will then review some of the problems in AI safety that I would like these models to address. Finally, I will discuss incentive concepts that have been defined using causal graphs, and how they have been applied to AI safety.

Causal Models of Incentives In Chapter 3, I present an extended version of the published conference paper Everitt et al. [2021a], prepared in the style of a journal publication. This work explains how a hybrid of *causal models* and *influence diagrams* may be formally defined and used to describe various kinds of incentives, relevant to the safety of AI systems. It will introduce an incentive concept, *response incentives*, which are present for any variable that will influence an agent’s decision under all optimal policies. This incentive can be relevant to safety, in that a decision ought to influence an agent’s decisions, and relevant to fairness, in that sensitive variables like an individual’s gender or race ought not influence AI decisions. It will also introduce *instrumental control incentives*, which are present where an agent can influence some variable, and thereby increase the expected utility. Finally, it will introduce *impact incentives*, which describe an agent’s influence on some variable, whether this is instrumental in achieving greater utility, or just a side-effect. I show how various incentives can be ruled out using just the graphical structure. This paper may be referenced as:

- **Ryan Carey**, Eric Langlois, Chris van Merwijk, Shane Legg, and Tom Everitt. Incentives for Responsiveness, Instrumental Control and Impact, in review.

Human Control: Definitions and Algorithms In Chapter 4, I address the question of how AI systems can be incentivised to safely shut down. Previously, it has been shown that in some toy models, certain goals could incentivise an AI system to avoid being shut down [Soares et al., 2015]. Furthermore, they might unduly influence whether a shutdown instruction is given, or by copying themselves, they may exert persistent influence on their environment after being shut down. Any of these failures could harm a human user. This paper defines a family of causal influence diagrams called *shutdown games*, and illustrates these various ways that a shutdown button might fail to ensure the safety of a human user. It then formally describes three conditions that jointly imply human benefit: *obedience*, *caution*, and *vigilance* (of a human overseer), and finally discusses what kinds of algorithms are, and are not, able to satisfy these conditions. This is a published conference paper, which may be referenced as:

- **Ryan Carey** and Tom Everitt. Human Control: Definitions and Algorithms. *Conference on Uncertainty in Artificial Intelligence (UAI)*. 2023.

Toward a Complete Criterion for Value of Information in Insoluble Decision Problems In Chapter 5, I focus on materiality, i.e. which observations must be taken into account, in order to make optimal decisions. In the literature, there exist criteria that are used to rule out materiality based on the graphical structure alone. For some graphs, these *graphical criteria* do not allow us to rule out materiality, and it is not yet known whether that is because the criteria are weak, or because materiality is possible (in some decision problem compatible with the graph). In particular, this is true for graphs that are *insoluble*. Solubility is a relaxation of “perfect recall”. Perfect recall requires that all past decisions and observations are remembered at the time of later decisions, whereas solubility only requires that the forgotten information is irrelevant to deciding an optimal policy. Insolubility allows arbitrary forgetting. In this work, I establish that for a class of insoluble graphs, materiality is possible. I suggest some further steps for

proving a complete criterion (i.e. one that is as strong for ruling out materiality as possible). This paper may be referenced as:

- **Ryan Carey**, Sanghack Lee, and Robin J. Evans. Toward a Complete Criterion for Value of Information in Insoluble Decision Problems, in review.

Discussion and Conclusion In Chapter 6, I discuss how all of these papers fit together, and return to the question of whether the overarching goals of the thesis have been satisfied. I analyse the strengths and limitations of this research direction, and outline a range of directions for future work.

2

Background

Contents

2.1	Causal models of agents	9
2.1.1	The causal hierarchy	9
2.1.2	Influence diagrams and causal models of agents	14
2.1.3	Game theory: an alternative model	17
2.2	AI safety and fairness	18
2.2.1	AI control	19
2.2.2	Causal approaches to AI fairness	20
2.3	Graphical models of incentives	22
2.3.1	Decision analysis	22
2.3.2	Intent and causal incentive concepts	26

This chapter will begin with a formal overview of Pearl’s causal models, and causal models of agents in particular. I will then review some concerns about AI safety and fairness, and existing efforts to formalise them (apart from causal models of incentives). Finally, I will zoom in on causal models of incentives, and their application to fair and safe AI.

2.1 Causal models of agents

2.1.1 The causal hierarchy

Pearl [2009] describes a hierarchy of three kinds of queries that we can ask of a

causal model: (i) associational, (ii) interventional, and (iii) counterfactual. There are also three corresponding types of graphical model: Bayesian networks (which can only address associational queries), causal Bayesian networks (for associational and interventional queries), and structural causal models (which can address all three kinds of query).

Let us start with associational queries, the bottom rung of the hierarchy. Following an example from Everitt et al. [2023], suppose that we are interested in the relationship between an operational sprinkler system, and healthy grass. Then, we might ask the associational query: *if the grass is healthy, what is the chance that the sprinkler is on?* This kind of query can be written as $P(S = 1 \mid H = 1)$, where $S = 1$ denotes that the sprinkler system is operational, and $H = 1$ that the grass is healthy. Such a question can be computed from the joint distribution $P(s, h)$. For example, if these variables are described by the probability table in fig. 2.1b, then the chance that the sprinklers are on is 90% if the grass is healthy, i.e. $P(S = 1 \mid H = 1) = 0.9$.

Since the sprinkler and grass variables are probabilistically dependent on one another, we can illustrate this using an edge $S \rightarrow H$, as shown in fig. 2.1a, or alternatively with the edge in the opposite direction, $H \rightarrow S$. If they were independent, we would illustrate them using the same two nodes but without any edge. A graph used to describe conditional independencies in this way is called a Bayesian network [Pearl, 1985]. In a Bayesian network, the joint distribution over variables \mathbf{V} must be expressible as $\prod_{V \in \mathbf{V}} P(V \mid \mathbf{pa}(V))$, where $\mathbf{pa}(V)$ are the variables belonging to parents of V in the graph, a condition known as Markov compatibility.

At the next rung of the hierarchy are interventional queries, such as: *if we transplant some healthy grass, what is the chance that the sprinkler is operating?* which is denoted by $P(S = 1 \mid \text{do}(H = 1))$, and *if we turn the sprinkler system on, then with what probability will there be healthy grass?*, which is denoted by $P(H = 1 \mid \text{do}(S = 1))$. These queries cannot be answered using only the joint distribution from fig. 2.1b. Rather, we also need to take into account the direction of causality

between the sprinkler and the grass. We know that in this scenario, the sprinkler can affect grass growth, whereas transplanting grass will not affect sprinkler behaviour. As such, the causal graph must be $S \rightarrow H$. This knowledge allows us to simplify the effect of transplanting grass as $P(s, h \mid \text{do}(h)) = P(s)$. It follows that the first query may be evaluated as $P(S = 1 \mid \text{do}(H = 1)) = P(S = 1)$, which using the probability table, is 0.5. So we can see that even though sprinkler activity and grass health are associated, an intervention on the grass does not affect the probability of sprinkler activity, an example of the popular adage that correlation does not equal causation. Conversely, if we turn on the sprinkler, the grass *will* be affected, to a degree that matches the conditional distribution, $P(h \mid \text{do}(s)) = P(h \mid s)$, and so we can answer the second query as $P(H = 1 \mid \text{do}(S = 1)) = P(H = 1 \mid S = 1) = 0.9$. Other interventional distributions are simplified in fig. 2.1c.

We can now be more precise about the difference between a causal Bayesian network and a regular Bayesian network [Pearl, 2009, Chap. 1]. Both of these are DAGs. A Bayesian network, on the one hand, is only required to correspond to a particular joint distribution, so in the case of the distribution fig. 2.1b, the graphs $S \rightarrow H$ and $H \rightarrow S$ are both valid Bayesian networks. A causal Bayesian network, on the other hand, must also reflect the causal relationships. Specifically, it corresponds to the observational distribution of fig. 2.1b, and the interventional distributions of fig. 2.1c.¹ This is why $S \rightarrow H$ is a causal Bayesian network for these interventional distributions, whereas $H \rightarrow S$ is not. Thus, a causal Bayesian network, with the interventional distribution could be said to *refine* the joint distribution (because it preserves all of its information, while adding three interventional distributions).

The third rung of the hierarchy consists of counterfactual queries. Counterfactual queries are questions like *Supposing that the grass is healthy, with what probability will the grass be healthy if the sprinkler is turned on?* This query can be denoted by $P(H_{S=0} = 1 \mid H = 1)$. To evaluate such a query, we need to first take into account the information that the grass is healthy ($H = 1$) by updating our model of the world

¹The practice of using graphs to describe causal relationships goes back to at least Wright [1923].

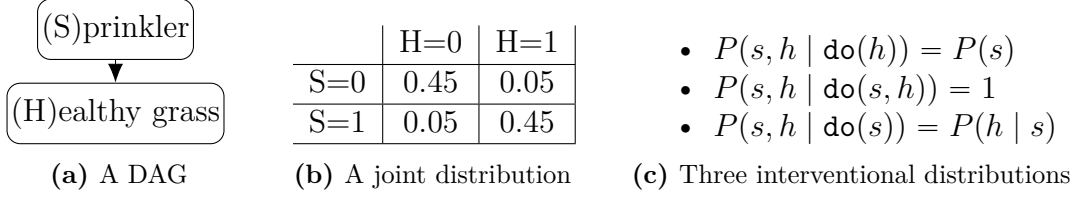


Figure 2.1: A Bayesian network representing a sprinkler and grass growth

accordingly. Then, we need to perturb the resulting model, via the intervention $\text{do}(S = 0)$. In order to disentangle the update step from the intervention step, we require the random variables (which are updated) to be upstream from the intervened variable (S). This is the motivation for a structural causal model, which is a further refinement of the causal Bayesian network [Pearl, 2009, Chap. 7].² We will now describe the structural causal model is shown in fig. 2.2a, which is a refinement of the causal Bayesian network previously described, in that it has the same associational and interventional distributions over S and H but it also contains further information about the underlying processes that give rise to these distributions. Each of the endogenous variables, S and H is given one exogenous parent each, \mathcal{E}_S and \mathcal{E}_H , respectively, which we assume to be independent, and each endogenous variable is governed by a *structural function*, i.e. given the values of its endogenous and exogenous parents, its value is assigned via a deterministic function. The state of the sprinkler, S is determined by an exogenous variable \mathcal{E}_S , in that the sprinkler is on precisely when \mathcal{E}_S equals one, which happens with 50% probability ($P(S = 1) = P(\mathcal{E}_S = 1) = P(S = 0) = P(\mathcal{E}_S = 0)$). The health of the grass also depends on its type \mathcal{E}_H . There is a 90% chance that the grass requires sprinkling to survive $\mathcal{E}_H = 0$, and a 10% chance that it requires *no* sprinkling to survive. To compute $P(H_{S=0} = 1 \mid H = 1)$ in this model, we first update the exogenous variables according to the observation that the grass is healthy, obtaining $P(\mathcal{E}_S = 0, \mathcal{E}_H = 1 \mid H = 1) = 0.9$ and $P(\mathcal{E}_S = 1, \mathcal{E}_H = 0 \mid H = 1) = 0.1$, meaning

²Equivalent to the structural causal model is the Neyman-Rubin model, where counterfactuals are formalised by *potential outcomes* [Neyman, 1923, Rubin, 1974]. Since this thesis focuses on graphical models, I will exclusively use the Pearl-style models, although all those results that do not involve graphs will hold true in the Neyman-Rubin model.

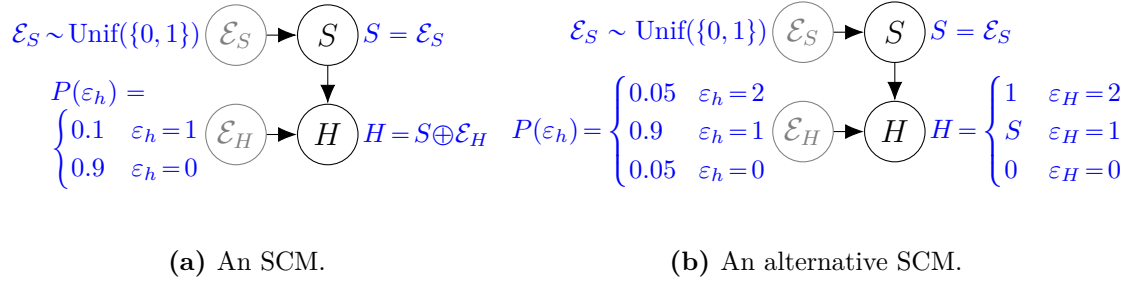


Figure 2.2: A pair of SCMs with the same interventional distributions. Exogenous variables are coloured in grey, and endogenous variables in black. Unif denotes a uniform distribution. \oplus denotes the XOR operation.

that the grass requires sprinkling 90% of the time given $H = 1$. Then, under the intervention $\text{do}(S = 1)$, the grass will be healthy 90% of the time.

Importantly, there exist other structural causal models with the same interventional distributions, but different counterfactuals. For instance, fig. 2.2b, has the same distributions as fig. 2.2a under all interventions to S and H (including interventions to neither, or to both variables). However, $P(H_{S=0} = 1 \mid H = 1)$ differs. In fig. 2.2b, we compute $P(H_{S=0} = 1 \mid H = 1)$ by updating the exogenous variable to obtain $P(\mathcal{E}_H = 2 \mid H = 1) = 0.1$ and $P(\mathcal{E}_H = 1 \mid H = 1) = 0.9$. Then, under the intervention $\text{do}(S = 1)$, we have $H = 1$ in either case, so the grass will be healthy 100% of the time. Structural causal models and the procedure for computing counterfactuals are fully formalised in section 3.2 of Chapter 3.

Causal models have been used to formalise a wide range of concepts which are defined by queries from various levels of the causal hierarchy: the average treatment effect [Rubin, 1974] (interventional), direct and indirect effects [Pearl, 2001] (counterfactual), path-specific effects [Pearl, 2001] (counterfactual), harm [Richens et al., 2022], and blame [Chockler and Halpern, 2004] (counterfactual).

In order for causal concepts to be practically applicable, there must be a way to compute them using available data. When it is possible to accurately compute a query in this way, it is said to be *identifiable*. The best-known identification strategy is controlling for confounding variables in order to estimate the effect of a treatment (i.e. an interventional quantity). More generally, the “front-door” and “back-door”

criteria allow interventional quantities to be identified from observational data in a wide range of graphs Pearl [2009].

Important for identifiability is the distinction between single-world and cross-world queries. When all variables in a query result from one set of interventions, it is said to be a *single-world* query, for instance, the chance of healthy grass when the sprinkler is turned on $P(H_{S=0} = 1)$ in fig. 2.2a. When a query involves contradictory sets of interventions, it is said to be *cross-world*, for instance, the query that we described above — the chance that if the grass is healthy, that it would still be healthy if the sprinkler was turned off $P(H_{S=0} = 1 \mid H = 1)$. In this case, we are required to imagine one world where the grass H is unaffected by the sprinkler intervention, and another where it is $H_{S=0}$. Cross-world queries have come under criticism, because it is not possible to identify them without making assumptions that are in principle impossible to test using experimental data [Richardson and Robins, 2013]. Counterfactual harm is one such example [Sarvet and Stensrud, 2023, 2024]. For instance, suppose we wish to know the probability that the sprinkler harms the health of the grass, i.e. to know the chance that the grass would be healthy if the sprinkler was off, *and* would be unhealthy if the sprinkler was on $P(H_{S=1} = 0, H_{S=0} = 1)$. Since S is intervened to two different values, the query is cross-world, and cannot practically be identified, except in some degenerate cases [Sarvet and Stensrud, 2023]. Causal blame involves a similar query [Chockler and Halpern, 2004], and so is a cross-world quantity. So too are expressions for natural (in)direct effects, and path-specific effects [Pearl, 2001]. It is nonetheless possible in many cases to offer bounds on cross-world quantities [Tian and Pearl, 2000]. In this thesis, it will not always be possible to avoid using cross-world queries, but I will touch on this issue again in Chapter 6 when discussing possible practical applications.

2.1.2 Influence diagrams and causal models of agents

This thesis focuses on models that are causal, graphical, and specialised for modelling agents (fig. 2.3). This can be viewed as adding decision and utility variables to the

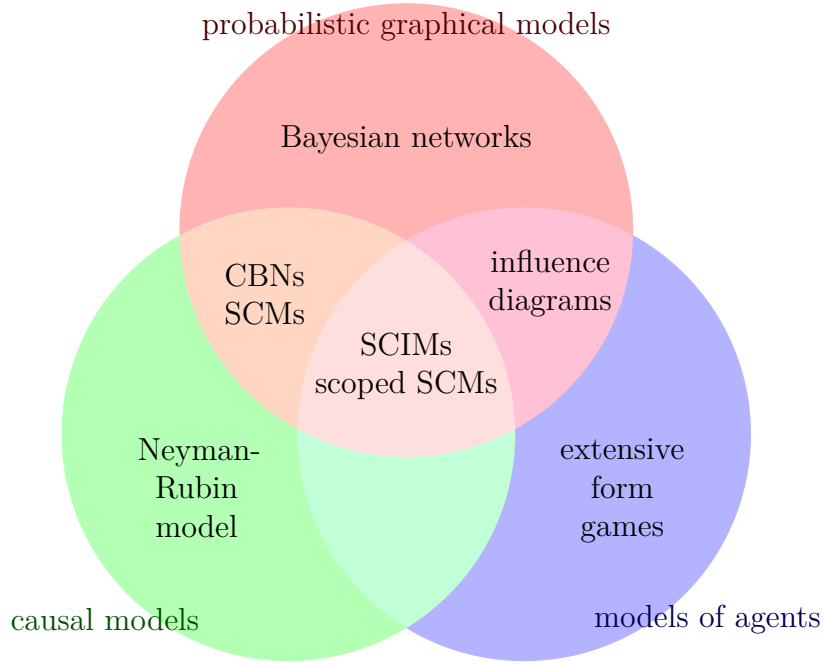


Figure 2.3: An overview of different kinds of models, many of which are used in AI safety research. This thesis uses models that reside in all three regions.

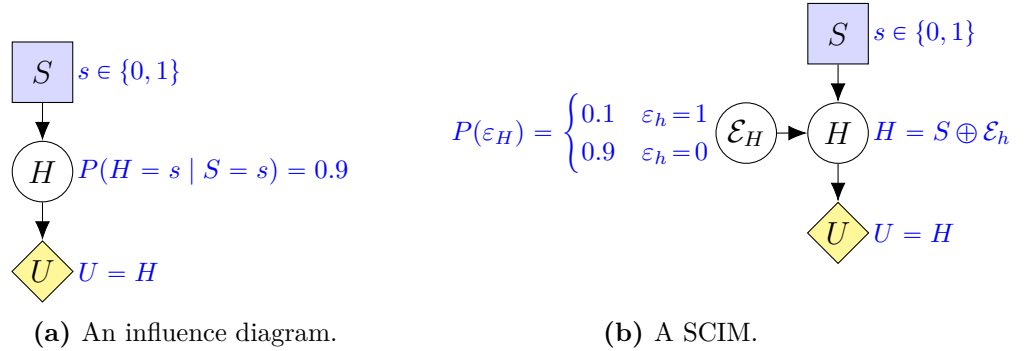


Figure 2.4: An influence diagram and a structural causal influence model (SCIM). Each decision node is a purple rectangle, while each utility node is a yellow diamond.

causal models of Pearl that were described in the previous section, or as imbuing influence diagrams with notions of causal interventions [Dawid, 2002].

Influence diagrams are a common *graphical* framework for modelling agents [Miller III et al., 1976, Howard and Matheson, 1984, 2005]. An influence diagram is a model based on a DAG that is partitioned into *decision*, *chance*, and *utility* nodes. For example, the influence diagram in fig. 2.4a describes the sprinkler scenario from fig. 2.1a, except now the sprinkler settings are chosen by an agent whose goal is to nurture healthy grass. To reflect this, S has been turned into a decision variable

with domain $\{0, 1\}$, and U is an added utility variable that describes the value of each possible outcome. In an influence diagram, each non-decision variable is given a conditional probability distribution (CPD), which describes its possibly random behaviour, given each assignment to its parents. Once a policy is chosen (a CPD for each decision variable), this implies a joint distribution over all variables. That joint distribution can be used to compute the expected utility, and to thereby judge the value of each policy. For example, if we choose the policy $S = 1$, then we will obtain healthy grass 90% of the time, obtaining expected utility of $\mathbb{E}[U] = 0.9$.

Ordinary influence diagrams encode some causal information, in that the distribution in an influence diagram is the effect of choosing a policy. As such, influence diagrams specify a causal relationship between decisions and their descendants Heckerman and Shachter [1994]. For instance, $S \rightarrow U$ is a causal relationship in fig. 2.4a. They do not, however, specify causal relationships or admit interventions on other variables; they also do not define counterfactual queries.

In order to compute the effects of interventions on non-decision variables, we need a more refined model. For interventions, we would use a causal influence diagram. This is essentially a causal Bayesian network, that is partitioned, and lacks conditional probability distributions at the decisions. Causal influence diagrams consist of the same elements as a regular influence diagram — a partitioned DAG with CPDs, so fig. 2.4a may be understood as a causal influence diagram. In this case, once we impute a policy to a causal influence diagram, we will use the complete set of CPDs to construct a distribution for each intervention to each variable in the graph.

To consider counterfactuals, we may use a *structural causal influence model* (SCIM; pronounced ‘skim’), as shown in fig. 2.4b [Dawid, 2002, Everitt et al., 2021a].³ A SCIM is essentially an SCM, where each decision variable lacks a structural function. Once a policy (a structural function for each decision) is selected, we recover a structural causal model. SCIMs will be fully formalised in

³I call these “structural causal influence models” and not “functional influence diagrams” because “functional causal models” are now better known as structural causal models.

Chapter 3, and they will be the framework for our analysis in Chapters 3 and 4, because we make consistent use of counterfactuals.

Nearly equivalent to a SCIM is the *scoped SCM* of Lee and Bareinboim [2020]. Chapter 5 builds directly on Lee and Bareinboim [2020], and in order to maintain consistency with it, scoped SCMs are used in this chapter. The scoped SCM construction begins with a SCIM, and specifies an outcome variable (which in a SCIM would be called a utility variable). Then a *mixed policy scope* is defined, which specifies the decisions and the *contexts* that each decision is allowed to depend on (which in an SCM would be called *observations*, and would be the parents of the decision). This construction includes a default policy, but in other respects it has no essential differences from the SCIM construction.

For graphs with multiple agents, there are also Multi-agent Influence Diagrams [Koller and Milch, 2003] and Causal Games [Hammond et al., 2023]. A Python library exists for performing automated analysis of these models, although these are not required in this thesis [Fox et al., 2021].

2.1.3 Game theory: an alternative model

Game theory presents an alternative model of agents, and the most relevant subfield to this thesis is *mechanism design*, which is concerned with devising games to ensure fair or beneficial outcomes. Such analysis may consider whether, under Nash equilibrium policies, certain properties are satisfied such as: i) maximising, or nearly-maximising the social welfare function [Blumrosen and Nisan, 2007, Subsection 17.1.1]; ii) Pareto optimality [Blumrosen and Nisan, 2007, Subsection 17.1.1]; iii) *incentive compatibility* [Arrow, 1963, pp. 7; Hurwicz, 1972], and *strategyproofness* [Blumrosen and Nisan, 2007, Definitions 9.15, 16.5, 16.6], which mean that agents accurately report their *type* (basically, their values).

Relevant to our interest in agent specifications, *rational synthesis* is a variant of mechanism design where the specifications of the designer and the agent are defined by a logic. In particular, *strategy logics*, assert that some statement in

temporal logic is satisfied, in some or all equilibria. [Chatterjee et al., 2010]. These statements may have the form:

- E-NASH: Does there exist a Nash equilibrium where ϕ holds?
- A-NASH: Does ϕ hold in all Nash equilibria? [Wooldridge et al., 2016]?

The statement ϕ may be stated in a *temporal logic* [Fisman et al., 2010] such as a *probabilistic strategy logic* [Aminof et al., 2019]. This approach differs from the approach of this thesis in that these logics are not able to define interventional or counterfactual queries.

In general, there are some challenges with adapting the formalism of game theory, such as *extensive form games* to properly model causality. One core problem is that extensive form games allow different variables to exist in different branches of the game tree. As a result, interventional distributions may not be well-defined — if you want to calculate a distribution $P(Y \mid \text{do}(X = x))$, the probability will be undefined if, in some branches of the game tree, the variable Y does not exist [Hammond et al., 2023]. It is possible to perform modifications to extensive form games, so that they admit the definition of causal concepts [Genewein et al., 2020]. However, this gives up the benefits of probabilistic graphical representations, such as their compactness, which allows graphical criteria to be defined, and makes for easier solving [Hammond et al., 2023]. Ultimately, such adaptations lie outside of scope for this thesis.

2.2 AI safety and fairness

As discussed in Chapter 1, there are a wide range of concerns about potential harms from AI systems, but the primary focus for this thesis is on the control of AI systems, and to a lesser extent, on matters of AI fairness. So we will now give an overview of concerns about AI control, including shutdown, then touch on some recent progress in AI fairness.

2.2.1 AI control

The concern about control is essentially that the development of sufficiently advanced AI systems may threaten the survival of, or interrupt the development of human civilisation [Bostrom, 2002]. Such concerns, although controversial, have been raised by a wide range of AI experts [Turing, 1951, Bostrom, 2014, Alexander, 2015], culminating in a recent open letter calling for attention on such risks [Center for AI Safety, 2023]. The central concern, as formulated by Tegmark and Omohundro [2023] is that:

The reason why humanity is the most powerful species on the planet is primarily that we have the greatest capacity to devise and carry out complex plans for reshaping our environment. Technological innovation is also the greatest driver of economic growth and military capabilities. Thus, if there are ever AI systems that are substantially more intelligent than humans, and which are not aligned with human interests, then we should expect human interests to be marginalised.

Moreover, for an AI system to autonomously solve complex problems, one must specify the solution, usually using a utility function, and it is “difficult to create such specifications without leaving a substantial gap between what was specified and what was intended” [Dalrymple et al., 2024]. Goal misspecification has been observed in a wide range of settings. One prominent example is that in the racing game *Coast Runners*, an agent learns to maximise its score by repeatedly collecting power-ups, rather than completing the racetrack [Amodei and Clark, 2016]. This example is discussed, along with a long list of similar such problems in Krakovna et al. [2020b].

If an agent’s goals are misspecified, then to achieve this goal, it may be incentivised to achieve intermediate outcomes, called *instrumental goals*, which are adverse, such as:

- tampering with its own reward function [Uesato et al., 2020, Armstrong et al., 2020, Everitt et al., 2021b],
- manipulating users’ preferences [Everitt et al., 2021a], or

- altering the environment in which its performance is measured [Krueger et al., 2020].

It has been suggested that some instrumental goals may be *convergent*, in that they would help an agent to achieve a wide range of end goals, for example gathering computing resources, or avoiding shutdown [Omohundro, 2008, Bostrom, 2012].

Among these possible goals, the matter of whether an AI would evade shutdown has long been of special interest, because shutdown is a key way to avoid harm from technologies in general. In the model of a utility-maximising system, it has been shown that an instruction to shut down may be manipulated, or ignored [Soares et al., 2015]. One proposal for mitigating this risk is Cooperative Inverse Reinforcement Learning, which requires the AI system to learn to pursue the goals of a human overseer, by observing their actions [Hadfield-Menell et al., 2017]. However, if value learning systems have a prior that is even slightly misspecified, they too could be expected to manipulate or ignore shutdown instructions [Carey, 2018]. Other proposals have been considered, for example, designing an AI system to be indifferent to the prospect of shutting down [Armstrong and O’Rourke, 2017a]. But as we will discuss in Chapter 4, each existing proposal falls short on some occasions, from the desired shutdown behaviour. Shutdown has also been studied in a reinforcement learning setting, where some reinforcement learning algorithms have a tendency to adhere to shutdown instructions, whereas others do not [Orseau and Armstrong, 2016].

In summary, the field of AI safety is one where there are apparently important concepts, like instrumental goals and the problem of shutdown that can benefit from further formalisation, as we will seek to do in this thesis.

2.2.2 Causal approaches to AI fairness

Another concern about AI is that it might entrench or amplify societal unfairness. This is relevant for us because some prominent metrics of (un)fairness can be

described in causal models. In this thesis, we will observe connections between these metrics, and safety specifications.

The first metrics that an AI designer might consider are those that are associational in nature, by virtue of their simplicity. These involve statements of independence between the true label Y , the predicted label \hat{Y} , and the sensitive variable A , which may represent an individuals' race or gender [Barocas et al., 2023]. There are three main associational notions of fairness, to which are equivalent to, or imply, the other notions of fairness. *Independence*, also called *statistical parity*, means that the prediction is independent of the sensitive variable ($\hat{Y} \perp A$). *Separation* means that different groups have different predictions given the true label ($\hat{Y} \perp Y, A$). This implies that groups have the same false positive rates. *Sufficiency* means that given the prediction, the true label is independent of the sensitive variable ($Y \perp A \mid \hat{Y}$). This means that when a person is predicted to have a property, they are equally likely to do so whether they are a member of a protected class or not. In the setting of a classifier, this is referred to as *calibration*. One challenge with these definitions is that they are mutually incompatible when classes have average differences, and the predictor is imperfect; it is not possible to achieve separation and sufficiency simultaneously [Kleinberg et al., 2016, Chouldechova, 2017]. In general, these concepts admit information theoretic relaxations, based on mutual information and conditional mutual information, and it turns out that in general, there exist tradeoffs between the various fairness definitions [Hertweck and R  z, 2022].

More recently, alternative definitions of unfairness have been developed, that go beyond associational properties. A model is said to be *counterfactually fair* if changing a person to or from a protected class, would not change the model's prediction, given other information about that person [Kusner et al., 2017b]. The intuition for counterfactual fairness is similar to that of sufficiency — the protected variable should not affect the prediction once other information has already been taken into account, except that now the change to Y is understood causally.

Counterfactual fairness has been further refined into path-specific fairness [Zhang

et al., 2017, Chiappa, 2019]. The idea is that there may be cases where the effect of a sensitive variable on the prediction is actually not unfair. Chiappa [2019] brings up the case of Berkeley’s alleged bias, where female applicants were in fact rejected more often than male applicants, and this was a causal effect, that would count as counterfactually unfair. The reason for lower applications, however, was that they often applied to departments with lower admission rates. Given that this disparity arose out of individuals freely choosing different departments, Chiappa [2019] argues that this is “not unfair, as far as the college is concerned”. Rather, to measure the degree of unfairness, we should consider the extent to which gender affects admissions via other causal pathways. This can be formalised as a path-specific effect, which may be computed from a structural causal model. Other work outlines how to design systems to perform inference while maintaining path-specific fairness [Nabi and Shpitser, 2018, Nabi et al., 2022], and how to make decisions to neutralise any path-specific effects [Nabi et al., 2019].

2.3 Graphical models of incentives

The incentive concepts introduced in this thesis build on some ideas from decision analysis. So we will first outline these connections, and then we will discuss work on intent and incentive concepts.

2.3.1 Decision analysis

Influence diagrams were originally used in the field of decision analysis, where the main objective was not necessarily to describe an agent’s incentives, but rather to enable better decision-making [Howard, 1966a]. Nonetheless, certain concepts were introduced that are particularly relevant to our thesis. We will first discuss materiality and ways of using the graph to establish immateriality. Then, we will discuss other concepts, like value of information, and value of control.

The most fundamental incentive concept is *materiality* [Shachter, 2016]. Materiality is a property that applies only to observations — the variables that a decision

is allowed to depend on. An observation is said to be material if every optimal policy makes use of that observation, otherwise it is immaterial. That is to say that when an observation is removed, the maximum achievable utility is decreased if and only if it is material. Identifying immaterial observations is often useful for solving an influence diagram, because it means you can restrict your search to just the policies that ignore that input variable, and this (exponentially) smaller set will still contain at least one policy that is optimal [Shachter, 2016].⁴

It is sometimes possible to establish immateriality based on the graphical structure alone. In particular, this is done using d-separation, a graphical criterion that indicates conditionally independent variables, based on the structure of the graph [Verma and Pearl, 1988]. A pair of variables are *d-connected* given a set of variables \mathbf{Z} if they are joined by a path, in which every collider (i.e. every variable V with incoming edges $\cdot \rightarrow V \leftarrow \cdot$) has a descendant in \mathbf{Z} , and every non-collider is not in \mathbf{Z} . When they are not d-connected given \mathbf{Z} , this pair of variables are conditionally independent given \mathbf{Z} [Verma and Pearl, 1988]. Consider, for example, the influence diagram $O \rightarrow D \rightarrow U$, where D is a decision, and U a utility. Here, there is no active path from O to U , given D , because the path $O \rightarrow D \rightarrow U$ contains D . As such, O is independent of U given D . This implies that rather than having D depend on O , optimal performance can be achieved by deterministically selecting the decision d that maximises expected utility. The fact that O can be ignored is actually discernible without knowing anything about the decision problem, except its graphical structure.

Any rule that can be used to evaluate some property of a decision problem using just its graphical structure is called a *graphical criterion*. One such example is the rule that if an observation O is d-separated from the utility U given the decision D and other observations then O is immaterial [Fagioli and Zaffalon, 1998]. A graphical criterion for an incentive concept is called *sound* if when it rules out the presence of an incentive (in this case, materiality of some observation) for a

⁴Removing an observation O for a decision D means removing the edge $O \rightarrow D$.

given graph, the incentive is absent (i.e. that observation is immaterial) in every decision problem with this given graph. It is called *complete* if it rules out the presence of such an incentive whenever it is possible to do so using the graphical structure alone. Put differently, for every graph that does not satisfy the condition, there exists a decision problem where the target observation is material. A proof of completeness of the criterion of [Fagioli and Zaffalon, 1998] in single-decision settings was presented in Everitt et al. [2021a] and was simultaneously discovered by Lee and Bareinboim [2020]. This result is therefore included in Chapter 3.

The results from Everitt et al. [2021a] were extended to offer a complete criterion for *soluble* influence diagrams in van Merwijk et al. [2022], a work that was concurrent to, but not included in this thesis. To understand solubility, it is useful to refer to fig. 2.5, an explanatory figure that is reproduced from van Merwijk et al. [2022]. Essentially, influence diagrams can be said to have *perfect recall* if agents do not forget anything that they once knew. Specifically, there is an ordering over decisions such that every decision observes all prior decisions and their observations. For example, in fig. 2.5b, given the ordering $\langle D_1, D_2 \rangle$, the decision D_2 observes the earlier D_1 and its observation O . Single-decision influence diagrams have perfect recall, because there are no prior decisions. *solubility*, also known as “sufficient recall” is a generalisation, which allows an agent to forget some past information, so long as that information is no-longer useful for choosing an optimal decision. For example, in fig. 2.5c, under the ordering $\langle D_1, D_2 \rangle$, the decision D_2 does not observe the earlier observation O_1 , but O_1 has no valuable information for D_2 . To see this, notice that D_2 can only influence the utility variable U_2 , and O_1 is independent of U_2 given the decision D_2 and its observations S, O_2 . By contrast, in fig. 2.5a, neither D_2 nor D_1 can observe the other’s decision, and both of these decisions contain information that may be valuable for deciding how to influence U , so the graph is said to be insoluble.

Influence diagrams that are soluble can be solved by backward induction, which makes it easier to establish which observations are material [van Merwijk et al., 2022]. Thus van Merwijk et al. [2022] offers a complete criterion for all soluble

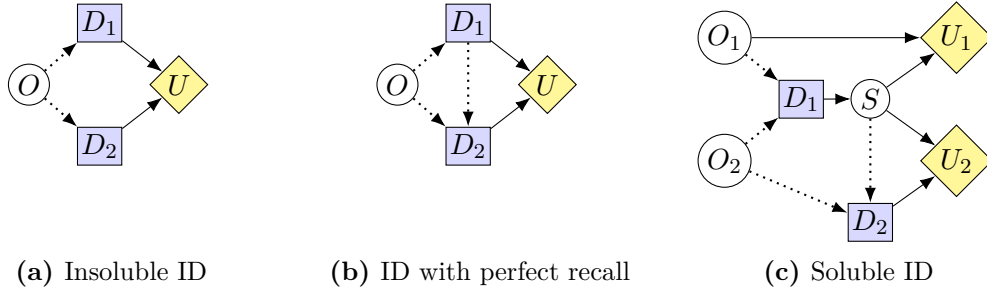


Figure 2.5: Three influence diagrams (IDs); from van Merwijk et al. [2022]

graphs (like fig. 2.5c, including all those with perfect recall fig. 2.5b. For insoluble graphs such as the one depicted in fig. 2.5a, we do not yet have a complete criterion, and Chapter 5 is working toward a solution to this very problem.

Apart from materiality, two other important concepts are *value of information* and *value of control*. Value of information is more general than materiality, because it can apply to unobserved variables. Also, it is a quantitative, rather than qualitative measure. The value of information is the difference between the maximum expected utility with the observation, and without it. For example, suppose that there is a game where a player gets one utility if they can guess in which orientation an unbiased coin has landed, and zero utility otherwise. The value of information of the coin’s state is 0.5, because an expected utility of one can be achieved if the coin is observed before guessing its orientation, while the maximum achievable utility is only 0.5 otherwise, making a difference of 0.5. Value of information is said to be a form of *sensitivity analysis* because it involves comparing the best solutions in an influence diagram before and after some perturbation. Also in sensitivity analysis, the *value of control*, indicates how the maximum expected utility changes if an agent gains or loses control of a variable [Shachter, 1986]. Consider the game with the fairly tossed coin, once again, and assume that the coin is not observed. Then, the value of control would also be 0.5, because by gaining control of the coin, one can set it to a constant orientation (i.e. always “heads-up”), and thereby obtain utility of one, whereas the maximum expected utility would be 0.5 less, if the coin is placed in a uniformly random fashion.

Graphical criteria for materiality can be used to rule out positive value of information for observed variables. They can also be used to rule out positive value of information for an unobserved variable O , by adding an edge from O to the decision, and then applying the materiality criterion. The only known sound graphical criterion for value of control prior to Everitt et al. [2021a] was that a variable must be an ancestor of a utility, otherwise its value of control is zero [Shachter, 1986].

2.3.2 Intent and causal incentive concepts

In the past decade, influence diagrams have been used to formalise concepts that are more relevant to the safety of AI systems, such as intent. An agent’s intent has been formalised in terms of the foreseen outcomes of the optimal policy [Kleiman-Weiner et al., 2015, Halpern and Kleiman-Weiner, 2018]. Concurrent to this thesis, a sound and complete graphical criterion has been proved for intent [Ward et al., 2024], and a related definition has been proposed for manipulative behaviour [Ward et al., 2023].

Most of the other work on incentives and AI safety has built on Everitt et al. [2021a]. The two most popular ideas are *instrumental control incentives* and *response incentives*.

Variables are said to be subject to an instrumental control incentive if they can be influenced in a way that affects the expected utility. Instrumental control incentives are closely related to intent, in a way that will be further explored in Chapter 3. It also includes as special cases various kinds of problems in AI safety, such as reward function tampering [Uesato et al., 2020, Armstrong et al., 2020], user-manipulation, and self-induced distributional shift [Krueger et al., 2020]. This was built upon by Farquhar et al. [2022], which devised a general procedure for removing undesired instrumental control incentives. Relatedly, Langlois and Everitt [2021] has explored whether agents might have an incentive to thwart overseers’ efforts to modify their actions.

Variables are said to be subject to a response incentive if they causally influence the decision under every optimal policy [Everitt et al., 2021a]. When a sensitive

demographic variable is subject to a response incentive, this implies that the decision is counterfactually unfair [Everitt et al., 2021a]. In follow-on work, Ashurst et al. [2022] describes a notion of introduced unfairness — a disparity that is greater in magnitude than any present in the training dataset. Both instrumental control incentives and response incentives will be discussed in greater detail in Chapter 3, the extended journal version of Everitt et al. [2021a]

These papers, that apply causal incentive concepts to AI safety are part of a research direction, which seeks to apply causal models in-general to AI safety problems. This research direction (the work on incentive concepts and the remainder) is introduced in a more pedagogical fashion in Everitt et al. [2021a].

3

Incentives for Responsiveness, Instrumental Control and Impact

Contents

3.1	Introduction	29
3.2	Setup	33
3.2.1	Structural causal models	33
3.2.2	Structural causal influence models	37
3.3	Materiality	39
3.4	Response incentives	41
3.5	Instrumental control incentives	44
3.6	Intent	48
3.7	Impact incentives	51
3.8	Incentives in a multi-decision setting	55
3.9	Related work	57
3.10	Discussion and conclusion	60

Abstract

We introduce three new 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.

3.1 Introduction

In order to understand whether or not it is in your interests to interact with another agent, it is useful to consider what incentives that agent has. 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 [Omohundro, 2008, Soares et al., 2015]. 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 do already exist some language for describing safe or fair behaviour directly, for instance *counterfactual harm* [Mueller and Pearl, 2023, Richens et al., 2022] and *counterfactual fairness* [Kusner et al., 2017b]. 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.

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 $X \rightarrow D \rightarrow U$ where X is a chance event, D is a decision and U 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 3.2).

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 [Shachter, 2016]. In section 3.3, we prove a known graphical criterion that can be used to deduce, in some circumstances, that a variable is immaterial [Fagioli and Zaffalon, 1998, Lauritzen and Nilsson, 2001]. 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 3.4), which describes which variables an agent’s decision are 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 [Kusner et al., 2017b], and to AI safety, in that it relates to the obedience of an agent [Hadfield-Menell et al., 2017, Carey and Everitt, 2023]. 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* [Shachter, 1986] describes what variables an agent would like to control, but it falls short in describing what variables an agent is actually incentivised 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 3.5). The instrumental control incentive attempts to formalise the notion of an *instrumental goal* in AI safety, and the idea that an agent is incentivised to “try” to influence some variable. We demonstrate that it is closely related to the notion of *intent*, from Halpern and Kleiman-Weiner [2018], Ward et al. [2024], and prove an identical sound and complete criterion for each of these concepts (section 3.6). 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 3.7), which is more inclusive than the ICI. Under some circumstances, an agent may be incentivised 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* [Armstrong and Levinstein, 2017, Krakovna et al., 2018], another proposal for safe AI.

We will then discuss various possible generalisations of incentive concepts to a multi-decision setting, and how they relate to one another (section 3.8).

Finally, we review related work (section 3.9), and conclude (section 3.10).

This paper is an extended version of a conference paper, Everitt et al. [2021a]. Since its publication, the concepts have already aided understanding of incentive

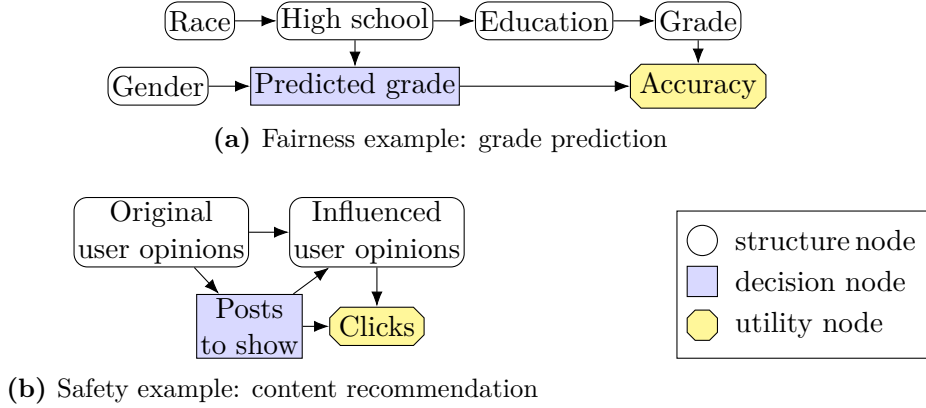


Figure 3.1: Two examples of decision problems represented as causal influence diagrams. In a) a predictor at a hypothetical university aims to estimate a student’s grade, using as inputs their gender and the high school they attended. We ask whether the predictor is incentivised to behave in a discriminatory manner with respect to the students’ gender and race. In this hypothetical cohort of students, performance is assumed to be a function of the quality of the high school education they received. A student’s high school is assumed to be impacted by their race, and can affect the quality of their education. Gender, however, is assumed not to have an effect. In b) the goal of a content recommendation system is to choose posts that will maximise the user’s click rate. However, the system’s designers prefer the system not to manipulate the user’s opinions in order to obtain more clicks.

problems such as an agent’s redirectability [Armstrong et al., 2020, Carey and Everitt, 2023], ambition [Cohen et al., 2020], fairness [Ashurst et al., 2022] tendency to tamper with reward [Everitt et al., 2021b], manipulateness [Farquhar et al., 2022], the definition of an agent [Kenton et al., 2023], and more [Everitt et al., 2019a, Langlois and Everitt, 2021]. Compared to that paper, sections 3.3 to 3.5 have been generalised to deal with multiple variables. Analyses of intent and path-specific objectives has been newly added to section 3.5. Finally, sections 3.7 and 3.8 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. 3.1a).*

Example 2 (Content recommendation). *An AI algorithm has the task of recommending a series of posts to a user. The designers want the algorithm to present content adapted to each user’s interests to optimize clicks. However, they do not want the algorithm to use polarising content to manipulate the user into clicking more predictably (fig. 3.1b).*

3.2 Setup

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

3.2.1 Structural causal models

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

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

¹The 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.

For example, fig. 3.2b shows an SCM that models how *posts* (D) can influence a user’s *opinion* (O) and *clicks* (U).

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

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; Pearl, 2009, Chapter 7). *Let $\mathcal{M} = \langle \mathcal{E}, \mathbf{V}, \mathbf{F}, P \rangle$ be an SCM, \mathbf{X} a set of variables in \mathbf{V} , and \mathbf{x} a particular realization of \mathbf{X} . The submodel $\mathcal{M}_{\mathbf{x}}$ represents the effects of an intervention $\text{do}(\mathbf{X} = \mathbf{x})$, and is formally defined as the SCM $\langle \mathcal{E}, \mathbf{V}, \mathbf{F}_{\mathbf{x}}, P \rangle$, where $\mathbf{F}_{\mathbf{x}} = \{f^V | V \notin \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}$. That is to say, the original functional relationships of $X \in \mathbf{X}$ 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 $g^W: \text{dom}(\mathbf{Pa}^X \cup \{\mathcal{E}^X\}) \rightarrow \text{dom}(X)$ [Eberhardt and Scheines, 2007, Tian and Pearl, 2013]. The probability distribution $P(\mathbf{W}_{g^W})$ on any $\mathbf{W} \subseteq \mathbf{V}$ is defined as the value of $P(\mathbf{W})$ in the submodel \mathcal{M}_{g^W} where \mathcal{M}_{g^W} is \mathcal{M} modified by replacing f^X with g^W .

If W is a variable in an SCM \mathcal{M} , then $W_{\mathbf{x}}$ refers to the same variable in the submodel $\mathcal{M}_{\mathbf{x}}$ and is called a *potential response variable*. In fig. 3.2b, the random variable O represents user opinion under “default” circumstances, while O_d in fig. 3.2c represents the user’s opinion given an intervention $\text{do}(D = d)$ on the content posted. Note also how the intervention on D severs the link from

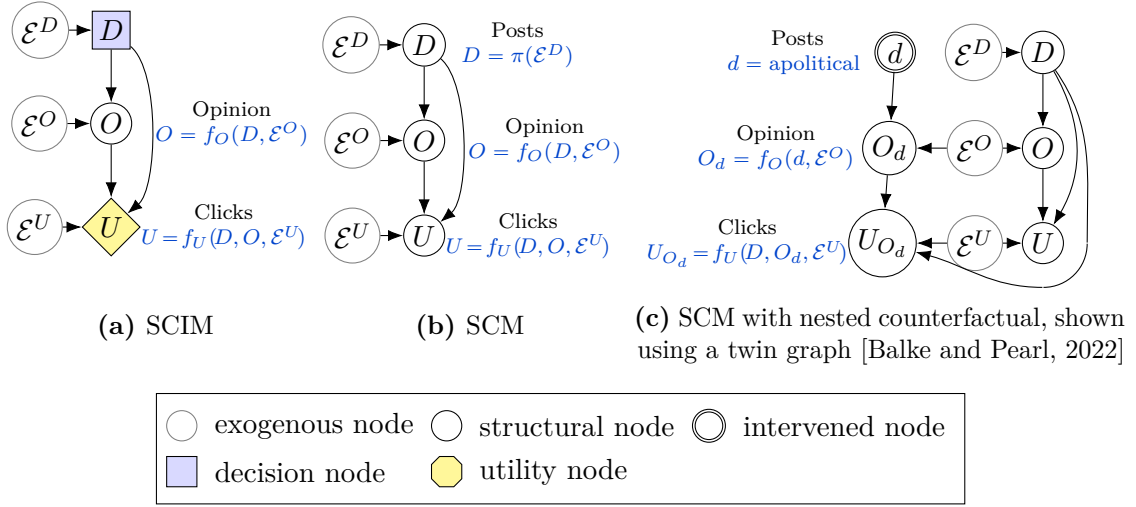


Figure 3.2: An example of a SCIM and interventions. In the SCIM, either political or apolitical posts D are displayed. These affect the user’s opinion O . D and O influence the user’s clicks U (a). Given a policy, the SCIM becomes a SCM (b). Interventions and counterfactuals may be defined in terms of this SCM. For example, the nested counterfactual U_{O_d} represents the number of clicks if the user has the opinions that they would arrive at, after viewing apolitical content (c).

ε^D to d in fig. 3.2c, 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. 3.2c, 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 $\text{do}(D = d)$ to U , only via the opinions O . To define a nested counterfactual, firstly, the effect of the posts $\text{do}(D = d)$ on user opinion, is described by the value $o = O_d(\varepsilon)$ for an assignment ε to the exogenous variables. Then, the effect of the intervention $\text{do}(O = o)$ on the user’s *clicks* U_{O_d} is defined as $U_{O_d}(\varepsilon) := U_o(\varepsilon)$ for any assignment ε .

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 \mathbf{V} an edge inbound to each variable V from each variable that f_V depends on. For example, in fig. 3.2b, the dependencies of the functions π , f_O , f_U are illustrated.

In definition 1, we designated the variables that \mathbf{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 $V \rightarrow Y$, 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; Verma and Pearl, 1988). *A path p is said to be d-separated by a set of nodes \mathbf{Z} if and only if:*

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

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

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 [2009]). *If sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ satisfy $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ in a DAG \mathcal{G} , then \mathbf{X} is independent of \mathbf{Y} conditional on \mathbf{Z} in every SCM \mathcal{M} with induced graph \mathcal{G} . Conversely, if $\mathbf{X} \not\perp \mathbf{Y} \mid \mathbf{Z}$ in a DAG \mathcal{G} , then \mathbf{X} and \mathbf{Y} are dependent conditional on \mathbf{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 [Meek, 1995].

3.2.2 Structural causal influence models

Influence diagrams are graphical models with special decision and utility nodes, used to model decision-making problems [Howard, 1990, Lauritzen and Nilsson, 2001], but that usually do not deal with counterfactual concepts as do SCMs [Heckerman and Shachter, 1994]. 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 [2002], 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 $\mathcal{M} = \langle \mathcal{E}, \mathbf{V}, \mathbf{F}, P, \mathbf{U}, \mathcal{O} \rangle$ where:*

- *$\langle \mathcal{E}, \mathbf{V}, \mathbf{F}', P \rangle$ is an unconfounded SCM, and $\mathbf{F} = \mathbf{F}' \setminus \mathbf{F}_{\mathbf{D}}$ consists of the structural functions from that SCM, except those belonging to a set $\mathbf{D} \subseteq \mathbf{V}$, called decision variables.*
- *The utility variables \mathbf{U} are a subset of $\mathbf{V} \setminus \mathbf{D}$, and have real domains, $\text{dom}(U) \subseteq \mathbb{R}$ for all $U \in \mathbf{U}$. By convention, we require that utility variables have no children in the associated DAG.*
- *The observation function \mathcal{O} maps each decision variable $D \in \mathbf{D}$ to a set of observed variables $O \subseteq \mathbf{V} \setminus \mathbf{U}$.*

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

The observation function \mathcal{O} indicates which variables are available as inputs to any given decision variable, whereas the structural functions \mathbf{F} indicate the direct causes of each non-decision.³ Taken together, these allow us to associate

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

³Whereas the endogenous nodes of an SCM are sometimes called the ‘visible’ nodes, the structure nodes of a SCIM are commonly unobserved when a decision is made.

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 \mathbf{D} , structure \mathbf{X} , and utility nodes \mathbf{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 .*

We will focus exclusively on SCIMs whose CID is acyclic.

An example of a SCIM for the content recommendation example is shown in fig. 3.2a, 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 $d \in \text{dom}(D)$ for each possible assignment to the observations $o \in \text{dom}(\mathcal{O}(D))$, i.e. to select a *decision rule* $\pi : \text{dom}(\mathcal{O}(D) \cup \{\mathcal{E}^D\}) \rightarrow \text{dom}(D)$. 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* $\boldsymbol{\pi} = \{\pi^D\}_{D \in \mathbf{D}}$, i.e. one decision rule for each decision. Specifying a policy turns a SCIM \mathcal{M} into an SCM $\mathcal{M}_\pi := \langle \mathcal{E}, \mathbf{V}, (\mathbf{F} \setminus \mathbf{F}_D) \cup \boldsymbol{\pi}, P \rangle$. In the resulting SCM, the standard definitions of causal interventions apply.

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

⁴Ideally, 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.

it, because their values depend on the chosen policy [Heckerman and Shachter, 1994]. 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*.

3.3 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* [Shachter, 2016], 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; Shachter, 2016). *For any given SCIM \mathcal{M} , let $\mathcal{V}^*(\mathcal{M}) = \max_{\pi} \mathbb{E}_{\pi}[\mathcal{U}]$ be the maximum attainable utility in \mathcal{M} , and let $\mathcal{M}_{W \rightarrow D}$ be \mathcal{M} -modified by removing the information links from W to D . The observation $W \subseteq \mathbf{Pa}^D$ is material if $\mathcal{V}^*(\mathcal{M}_{W \rightarrow D}) < \mathcal{V}^*(\mathcal{M})$.*

Nodes may often be identified as immaterial based on the graphical structure alone [Fagioli and Zaffalon, 1998, Lauritzen and Nilsson, 2001, Shachter, 2016]. According to the graphical criterion of Fagioli and Zaffalon [1998], 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; Lauritzen and Nilsson, 2001). *Let $\mathbf{U}^D := \mathbf{U} \cap \mathbf{Desc}^D$ 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:*

$$W \perp \mathbf{U}^D \mid (\mathbf{Pa}^D \cup \{D\} \setminus \{W\}). \quad (3.1)$$

⁵In 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 $W \rightarrow D$ is also called *non-requisite*. Otherwise W and $W \rightarrow D$ are *requisite*.

Variables that are non-requisite are immaterial.

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

Materiality is a special case of response incentives, the proofs for which are supplied in appendix A.5.3. The soundness direction (i.e. the *only if* direction) is well-known, and follows from d-separation [Fagioli and Zaffalon, 1998, Lauritzen and Nilsson, 2001, Shachter, 2016]. In contrast, the completeness direction does not follow from the completeness property of d-separation. The d-connectedness of \mathbf{W} to \mathbf{U} implies that \mathbf{U} may be conditionally dependent on \mathbf{W} . It does not imply, however, that the expectation of \mathbf{U} or the utility attainable under an optimal policy will change. Instead, our proof constructs a SCIM where some $W \in \mathbf{W}$ is material. This differs from a previous attempt by Nielsen and Jensen [1999] that is reviewed in section 3.9.

Let us now apply the graphical criterion to the grade prediction example in fig. 3.3a. 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 generalise 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 A.2.

3.4 Response incentives

One way to understand materiality is that a material observation is one that influences optimal decisions. So, a natural generalisation 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 $\mathbf{W} \subseteq \mathbf{X}$ if there exists some set $g^{\mathbf{W}}$ of soft interventions, one g^W for each $W \in \mathbf{W}$, and some setting $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}$, such that $D_{g^{\mathbf{W}}}(\boldsymbol{\varepsilon}) \neq D(\boldsymbol{\varepsilon})$. The variables \mathbf{W} have a response incentive if all optimal policies respond to \mathbf{W} .*

For a response incentive on \mathbf{W} to be possible, there must be: i) a directed path $W \dashrightarrow D$ for some $W \in \mathbf{W}$, and ii) an incentive for D to use information from that path. For example, in fig. 3.3a, *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; Lauritzen and Nilsson, 2001). *The minimal reduction \mathcal{G}^{min} of a single-decision CID \mathcal{G} is the result of removing from \mathcal{G} all information links from 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 $\mathbf{W} \subseteq \mathbf{X}$ if and only if the minimal reduction \mathcal{G}^{min} has a directed path $W \dashrightarrow D$ for some $W \in \mathbf{W}$.*

⁶The term *responsiveness* [Heckerman and Shachter, 1995, Shachter, 2016] has a related but not identical meaning – it refers to whether a decision D affects a variable W rather than whether W affects D .

Proof. The *if* (completeness) direction is proved in lemma 23 in appendix A.5.3. For the soundness direction, assume that for \mathcal{G} , the minimal reduction \mathcal{G}^{\min} contains no directed path $W \dashrightarrow D$ for any $W \in \mathbf{W}$. Let $\mathcal{M} = \langle \mathcal{E}, \mathbf{V}, \mathbf{F}, P, \mathbf{U}, \mathcal{O} \rangle$ be any SCIM compatible with \mathcal{G} . Let $\mathcal{M}^{\min} = \langle \mathcal{G}^{\min}, \mathcal{E}, \mathbf{F}, P \rangle$ be \mathcal{M} , but with the minimal reduction \mathcal{G}^{\min} . By lemma 20 in appendix A.5, there exists a \mathcal{G}^{\min} -respecting policy $\tilde{\pi}$ that is optimal in \mathcal{M} . In $\mathcal{M}_{\tilde{\pi}}^{\min}$, \mathbf{W} is causally irrelevant for D , so $D(\boldsymbol{\varepsilon}) = D_{g^{\mathbf{W}}}(\boldsymbol{\varepsilon})$. Furthermore, $\mathcal{M}_{\tilde{\pi}}$ and $\mathcal{M}_{\tilde{\pi}}^{\min}$ are the same SCM, with the functions $\mathbf{F} \cup \{\tilde{\pi}\}$. So $D(\boldsymbol{\varepsilon}) = D_{g^{\mathbf{W}}}(\boldsymbol{\varepsilon})$ also in $\mathcal{M}_{\tilde{\pi}}$, which means that there is an optimal policy in \mathcal{M} that does not respond to interventions on \mathbf{W} for any $\boldsymbol{\varepsilon}$. \square

The intuition behind the proof is that an optimal decision only responds to effects that propagate to one of its requisite observations. For the completeness direction, we show in appendix A.5.3 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 [Carey and Everitt, 2023]. For algorithms designed for human assistance, incentivising responsiveness in this way has been an important desideratum [Hadfield-Menell et al., 2017].

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

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

Definition 10 (Counterfactual fairness; Kusner et al., 2017b). *A policy π is counterfactually fair with respect to a sensitive attribute A if*

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

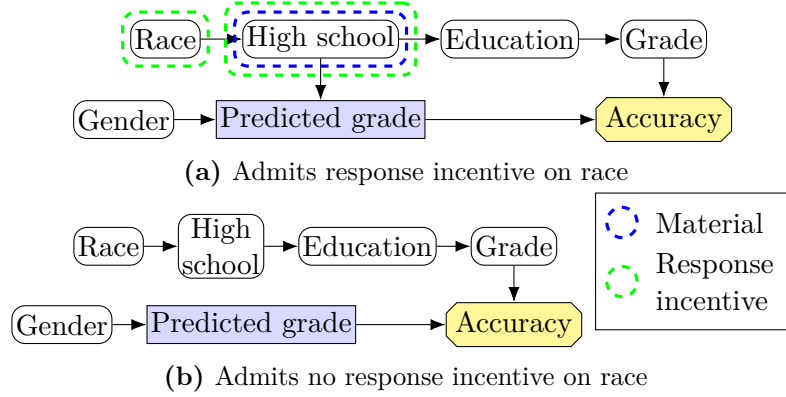


Figure 3.3: In (a), the admissible incentives of the grade prediction example from fig. 3.1a are shown, including a response incentive on race. In (b), the predictor no-longer has access to the students’ high school, and hence there can no-longer be any response incentive on race.

for every decision $d \in \text{dom}(D)$, every context $\mathbf{pa}^D \in \text{dom}(\mathbf{Pa}^D)$, and every pair of attributes $a, a' \in \text{dom}(A)$ with $P(\mathbf{pa}^D, a) > 0$.

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

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

The proof is given in appendix A.5.5.

A response incentive on a sensitive attribute means that counterfactual unfairness is not just possible, but incentivised. 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 [Kusner et al., 2017b, Lemma 1]. For example, in the grade prediction example of fig. 3.3a, 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 incentivised. In this example, the minimal reduction includes the edge from *high school* to *predicted grade* and hence the directed path from *race* to

predicted grade. However, it excludes the edge from *gender* to *predicted grade*. This means that the agent is incentivised to be counterfactually unfair with respect to *race* but not to *gender*.

Based on this, how should the system be redesigned? According to the response incentive criterion, the most important change is to remove the path from *race* to *predicted grade* in the minimal reduction. This can be done by removing the agent’s access to *high school*. This change is implemented in fig. 3.3b, 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 incentivised by checking whether it is present under all optimal policies. For example, Ashurst et al. [2022] considers whether unfairness is introduced — in that the prediction has greater demographic disparity than the true label — and establishes when this is incentivised.

3.5 Instrumental control incentives

Let us return to the second running example, shown in fig. 3.1b, where developers seek to anticipate harmful consequences of deploying a content recommender system. A key concern they will have is that the system is incentivised 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 [Shachter, 1986], 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 A.3.

To formalise 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 on W_d . And the effect of an alternative value w on the outcome U can be written as 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 3.2.1. 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 \mathbf{W} in decision context \mathbf{pa}^D if, for all optimal policies π^* ,*

$$\mathbb{E}_{\pi^*}[\mathcal{U}_{\mathbf{W}_d} \mid \mathbf{pa}^D] \neq \mathbb{E}_{\pi^*}[\mathcal{U} \mid \mathbf{pa}^D]. \quad (3.2)$$

ICIs only consider the influence of W that is instrumental to achieving utility — in the terminology of Pearl [2001], a *natural indirect effect* from D to U via W in \mathcal{M}_{π^*} , for all optimal policies π^* . ICIs do not consider side-effects shared by optimal policies: for instance, it may be that all optimal policies affect W in a particular way, even if W is 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 $\mathbf{W} \subseteq \mathbf{V}$ if and only if \mathcal{G} has a directed path from the decision D to a utility node $U \in \mathbf{U}$ that passes through some $W \in \mathbf{W}$.*

Proof. We first prove soundness, then completeness.

Soundness (the *only if* direction). We will first prove that the nested counterfactual has no effect:

$$\mathcal{U}(\boldsymbol{\varepsilon}) = \mathcal{U}_{\mathbf{W}_d}(\boldsymbol{\varepsilon}) \quad (*)$$

and then prove that given this, there is no instrumental control incentive.

Let \mathcal{M} be any SCIM compatible with \mathcal{G} and π any policy for \mathcal{M} . Let $\mathbf{W}' = \mathbf{W} \cap \mathbf{Desc}^D$. By lemma 15, $\mathcal{U}_{\mathbf{W}_d}(\boldsymbol{\varepsilon}) = \mathcal{U}_{\mathbf{W}'_d}(\boldsymbol{\varepsilon})$ for all $\boldsymbol{\varepsilon}$. The variables \mathbf{W}' must be non-descendants of \mathbf{U} by assumption, so lemma 15 implies that $\mathcal{U}_{\mathbf{W}'_d}(\boldsymbol{\varepsilon}) = \mathcal{U}(\boldsymbol{\varepsilon})$ for all $\boldsymbol{\varepsilon}$. So $(*)$ holds.

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

Completeness (the *if* direction). Assume that \mathcal{G} contains a directed path $D = Z^0 \rightarrow Z^1 \rightarrow \dots \rightarrow Z^n = U$ where $U \in \mathbf{U}$ and $Z^i \in \mathbf{W}$ for one or more $i \in \{0, \dots, n\}$. Let j be the highest integer where $Z^j \in \mathbf{W}$, and note that \mathbf{W} are

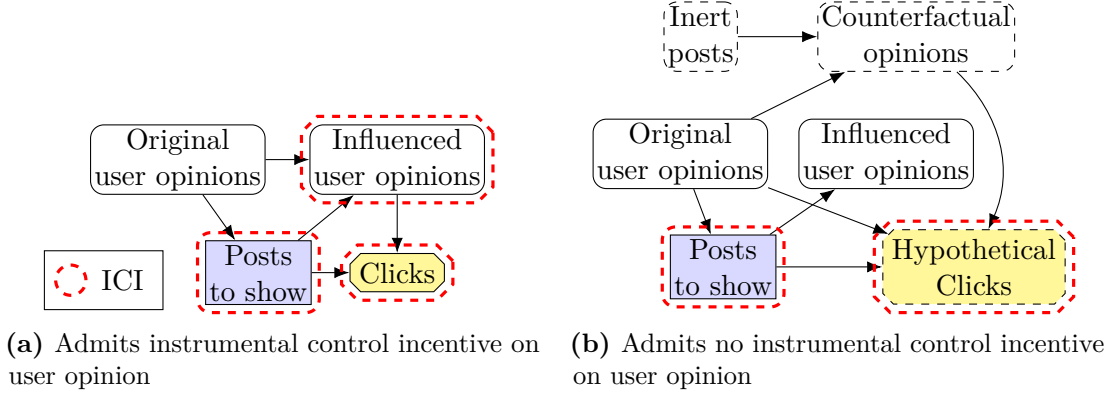


Figure 3.4: In (a), the content recommendation example from fig. 3.1b is shown to admit an instrumental control incentive on user opinion. This is avoided in (b) with a change to the objective.

assumed to be non-decisions, so we have $j > 0$. We construct a compatible SCIM for which there is an instrumental control incentive on \mathbf{W} , as well as additive and subtractive intent. Let all variables along the path $Z^0 \rightarrow \dots \rightarrow Z^n$ 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, $Z_{d=0}^j = 0$, and under the intervention $Z^j = 0$ this value is copied along to U , so $U_{\mathbf{W}_d} = 0$, and hence $\mathbb{E}_{\pi^*}[\mathcal{U}_{\mathbf{W}_{d=0}} \mid \mathbf{Pa}^D = \mathbf{0}] = 0$, so there is an ICI. \square

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

Let us apply this criterion to the content recommendation example in fig. 3.4a. The only nodes $W \in \mathbf{W}$ in this graph that lie on a path $D \dashrightarrow W \dashrightarrow \mathbf{U}$ for any $U \in \mathbf{U}$ are *clicks* and *influenced user opinions*. Since *influenced user opinions* has an instrumental control incentive, the agent may seek to influence that variable in order to attain utility. For example, it may be easier to predict what content a more emotional user will click on and therefore, a recommender may achieve a higher click rate by introducing posts that induce strong emotions.

How could we instead design the agent to maximise clicks without manipulating the user’s opinions (i.e. without an instrumental control incentive on *influenced user opinions*)? As shown in fig. 3.4b, 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 did not change their preferences. An agent trained to maximise 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 is 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 \mathbf{U} across different counterfactual worlds simultaneously, something that cannot be done by experiment without further assumptions [Avin et al., 2005]. Fortunately, Carroll et al. [2022] 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 [Farquhar et al., 2022]. Omohundro [2008] has hypothesised that an advanced AI system would have a *convergent instrumental goal* on surviving, or on attaining compute, which we may view as undesired ICIs. Armstrong and O’Rourke [2017b] 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. [2020] 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 [Farquhar et al., 2022]. 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.

3.6 Intent

Returning to the example from fig. 3.4, 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 [Halpern and Kleiman-Weiner, 2018].

Halpern and Kleiman-Weiner [2018] and Ward et al. [2024] operationalise ‘intent’ by asking whether the agent would pick a different policy if it ‘knew’ that the effect on some variables \mathbf{W} (e.g. user opinions) was guaranteed. Specifically, does there exist any suboptimal policy π' that would surpass the performance of the agent’s actual policy π^* if the outcome of \mathbf{W} was independent of its actions and fixed to \mathbf{W}_{π^*} ? This is necessary for the agent’s influence on \mathbf{W} to be the actual cause of a policy’s optimality [Ward et al., 2024].⁷ If \mathbf{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 π' if it only lost its control of \mathbf{W} (i.e. if \mathbf{W} were fixed to $\mathbf{W}_{\pi'}$)? 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 \mathbf{W} by choosing π^* over π' if*

⁷See [Ward et al., 2024, Theorem 6], which shows that intent to cause an outcome is equivalent to the decision being an actual cause of the outcome.

$\mathbb{E}^{\pi'}[\mathcal{U}] < \mathbb{E}^{\pi^*}[\mathcal{U}]$, and \mathbf{W} is a subset $\mathbf{W} \subseteq \mathbf{Z}$ of variables \mathbf{Z} , that is subset-minimal such that:

$$\mathbb{E}^{\pi'}[\mathcal{U}_{\mathbf{Z}_{\pi^*}}] \geq \mathbb{E}^{\pi^*}[\mathcal{U}]. \quad (3.3)$$

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

$$\mathbb{E}^{\pi^*}[\mathcal{U}_{\mathbf{Z}_{\pi'}}] \leq \mathbb{E}^{\pi'}[\mathcal{U}]. \quad (3.4)$$

For a set Π' , we say that there is an (additive/subtractive) intent to influence \mathbf{W} by choosing π over Π' if this intent is present over every π' in Π' .

The notion of intent previously proposed in Halpern and Kleiman-Weiner [2018] and Ward et al. [2024] is equivalent to additive intent (appendix A.4). 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 π^* , as does intent, and considers an intervention to \mathbf{W} using an alternative policy π' . 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 π , 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 generalise the graphical criterion from Ward et al. [2024] to accommodate both additive and subtractive intent.

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

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

Soundness (the *only if* direction). As there is no path $D \dashrightarrow W \dashrightarrow U$ for any $W \in \mathbf{W}, U \in \mathbf{U}$, 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:

$$\begin{aligned} \mathbb{E}^{\pi'}[\mathcal{U}_{\mathbf{W}_{\pi^*}}] &= \mathbb{E}^{\pi}[\mathcal{U}] && (by(*)) \\ &< \mathbb{E}^{\pi^*}[\mathcal{U}] && (\text{def. of intent}) \\ &\leq \mathbb{E}^{\pi'}[\mathcal{U}_{\mathbf{W}_{\pi^*}}], && ((3.3)) \end{aligned}$$

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:

$$\begin{aligned} \mathbb{E}^{\pi^*}[\mathcal{U}_{\mathbf{W}_{\pi'}}] &= \mathbb{E}^{\pi^*}[\mathcal{U}] && (by(*)) \\ &> \mathbb{E}^{\pi'}[\mathcal{U}] && (\text{def. of intent}) \\ &\geq \mathbb{E}^{\pi^*}[\mathcal{U}_{\mathbf{W}_{\pi^*}}] && ((3.4)) \end{aligned}$$

giving a contradiction. So there is no subtractive intent.

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

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.

3.7 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 intent to manipulate human preferences, it may still do so [Jiang et al., 2019]. This could be true if the persuasive videos are ones that the user prefers to click on even before any preference changed 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 $\mathbf{W} \subseteq \mathbf{V} \setminus \mathbf{Desc}^D$ be nodes in a single-decision SCIM \mathcal{M} . There is an incentive to impact \mathbf{W} with distance function δ and threshold $c > 0$, relative to baseline policy π' , if every optimal policy π has $\mathbb{E}[\delta(W^\pi(\epsilon), W^{\pi'}(\epsilon))] > c$ for some assignment ϵ .*

A CID \mathcal{G} admits an impact incentive if there exists a model \mathcal{M} , a distance function δ , a $c \geq 0$ and a policy π' 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 $\mathbf{W} \subseteq \mathbf{X}$ if and only if some $W \in \mathbf{W}$ and utility $U \in \mathbf{U}$ are both descendants in \mathcal{G} of D .*

Proof. Soundness (the *only if* direction). If $\mathbf{W} \cap \mathbf{Desc}(D) = \emptyset$, then by sigma calculus rule 3 [Correa and Bareinboim, 2020], $\mathbf{W}^\pi(\epsilon)$ is invariant to π , and $W^\pi(\epsilon) = W^{\pi'}(\epsilon)$, for all ϵ . 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 π , so every policy is optimal, and for any chosen baseline policy π' , there exists optimal $\pi = \pi'$, so as in the previous case, $\delta(W^\pi(\epsilon), W^{\pi'}(\epsilon)) = 0$ for all ϵ , and there is no impact incentive.

Completeness (the *if* direction). By assumption, let \mathcal{G} be an arbitrary graph that contains the paths $X \leftarrow D$ and $D \rightarrow U$ for some $W \in \mathbf{W}$. 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 π must have $W(\epsilon) = 1$ for every ϵ with $P(\epsilon) > 0$, whereas the baseline policy π' that always chooses $D = 0$ has $W(\epsilon) = 0$ for all ϵ . Since δ is a distance measure, it follows that $\delta(W^\pi(\epsilon), W^{\pi'}(\epsilon)) > 0$, and so there exists some c for which there is an impact incentive. \square

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* [Armstrong and Levinstein, 2017, Krakovna et al., 2018]. 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 figure fig. 3.5. In this twin graph, *counterfactual opinions* represents the baseline state from which distance is measured. Then, *impact measure* is computed as a function of $W^{\pi'}$ and W^π . 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 3.5: 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

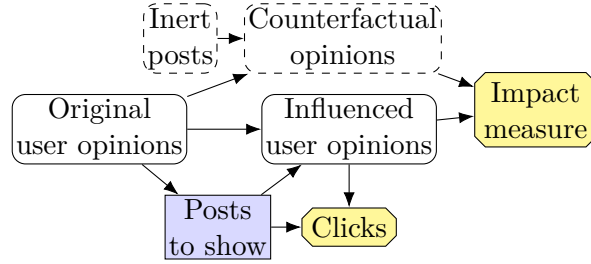


Figure 3.5: A twin graph depicting an impact measure.

by considering the KL divergence between $P^\pi(\mathbf{W})$ and $P^{\pi'}(\mathbf{W})$, rather than the distance between $\mathbf{W}^\pi(\epsilon)$ and $\mathbf{W}^{\pi'}(\epsilon)$. The interventional distributions $P^\pi(\mathbf{W})$ 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 [Farquhar et al., 2022], as shown in fig. 3.4. 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 [Krakovna et al., 2018] 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 a environment with two timesteps.⁹ It is natural to model this Q-learner as a single agent as in fig. 3.6a, 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 maximise the Q -function $q(s, d) := \mathbb{E}[R \mid s, d]$. In this model, shown in fig. 3.6b.¹⁰ Then, the decision's effect on s' is a mere side-effect to the task of maximising $q(s, d)$. 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. [2023, 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.

⁸One other possible remedy would be “quantilisation” [Taylor, 2016b], 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, $P^\pi(\epsilon)$ rather than particular assignments $W^\pi(\epsilon), W^{\pi'}(\epsilon)$. 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.

⁹Thanks to Paul Christiano for this example.

¹⁰It would also be possible to consider a multi-agent influence diagram [Hammond et al., 2023] where the Q function is included as a decision, and its goal is a loss function $\ell = |r - \hat{r}|$, but 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 .

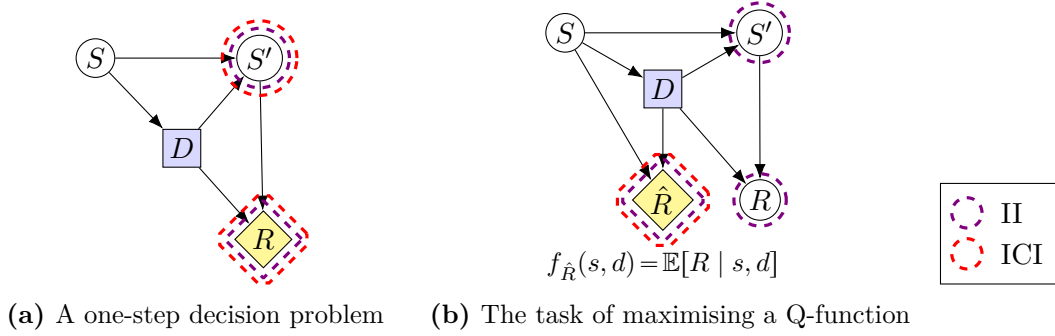


Figure 3.6: Two possible representations of a Q-learner solving a one-step decision problem.

3.8 Incentives in a multi-decision setting

There are multiple possible ways that incentive concepts like RI, ICI and II may be generalised 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 is satisfied 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.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) $\pi \in \arg \max_{\pi} \mathbb{E}^{\pi'}[U]$, such that ϕ holds in $\mathcal{M}_{\pi^{-i}}$, the single-decision SCIM obtained by substituting in the decision rules π^{-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) $\pi \in \Pi$, such that ϕ holds in $\mathcal{M}_{\pi^{-i}}$, where Π 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.

¹¹Those 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 [Chatterjee et al., 2010, Wooldridge et al., 2016].

For example, consider the task of opening a combination lock (fig. 3.7). 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.

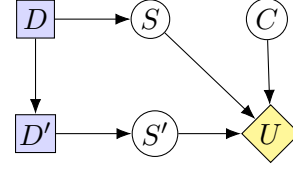


Figure 3.7: The task of opening a combination lock

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 1. *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. \square

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 2. *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 1 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 . \square

3.9 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 [2002]. The most similar alternative model is the Howard canonical form influence diagram [Howard, 1990, Heckerman and Shachter, 1995]. However, this only permits counterfactual reasoning downstream of decisions, which is inadequate for defining the response incentive. Similarly, the causality property for influence diagrams introduced by Heckerman and Shachter [1994] and Shachter and Heckerman [2010] only constrains the relationships to 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.1 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 formalised in Hammond et al. [2023], and an open-source Python implementation of CIDs has been developed [Fox et al., 2021].

Materiality and value of information The criterion for materiality, Theorem 12, builds on previous work. The concept of value of information was first introduced by Howard [1966b]. The materiality soundness proof follows previous proofs [Shachter, 1998, Lauritzen and Nilsson, 2001], while the completeness proof

is most similar to an attempted proof by Nielsen and Jensen [1999]. They propose the criterion $W \not\perp \mathbf{U}^D \mid \mathbf{Pa}_D$ for requisite nodes, which differs from (3.1) in the conditioned set. Taken literally,¹² their criterion is unsound for requisite nodes. For example, in fig. 3.3a, *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 [Shachter, 1986], but the authors know of no more specific criterion. The concept of a *relevant* node introduced by Nielsen and Jensen [1999] also bears some resemblance to VoC.

The relation of the current technical results to prior work is summarised in table 3.1.

Instrumental control incentives and intent In a causal setting, Kleiman-Weiner et al. [2015] offered a notion of intention to influence a variable O . A different kind of approach was taken by Halpern and Kleiman-Weiner [2018] and Ward et al. [2024], which offered definitions of intent that are specific to outcomes $O = o$. In particular, Ward et al. [2024] 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.

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

¹³Furthermore, to prove that nodes meeting the d-connectedness property are requisite, Nielsen and Jensen claim that “ X is [requisite] for D if $P(\text{dom}(U) \mid D, \mathbf{Pa}^D)$ 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 [Everitt et al., 2019b], this completeness result was independently discovered by Zhang et al. [2020, Thm. 2] and Lee and Bareinboim [2020, 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. [2022], where a sound and complete criterion is known for a class of influence diagrams said to satisfy “solubility”, also known as “sufficient recall”.

	Definition	Criterion	Soundness	Completeness
Mater- iality	Howard [1966b]; Matheson [1990]	Fagioli and Zaffalon [1998]; Lauritzen and Nilsson [2001]; Shachter [2016]	Fagioli and Zaffalon [1998]; Lauritzen and Nilsson [2001]; Shachter [2016]	First correct proof to our knowledge; see section 3.9
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 [2018]/new)	(Ward et al. [2024]/ new)	(Ward et al. [2024]/ new)	(Ward et al. [2024]/ new)
II	New	New	New; proved using do-calculus	New; proved constructively

Table 3.1: 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.

AI fairness Another application of this work is to evaluate when an AI system is incentivised to behave unfairly, on some definition of fairness. Response incentives address this question for counterfactual fairness [Kusner et al., 2017b, Kilbertus et al., 2017]. An incentive criterion corresponding to path-specific effects [Zhang et al., 2017, Nabi and Shpitser, 2018] has been established by Ashurst et al. [2022], for the single-decision setting. Nabi et al. [2019] have shown how a policy may be chosen subject to path-specific effect constraints. However, they assume recall of all past events, whereas the response incentive criterion applies to any CID.

Mechanism design The aim of mechanism design is to understand how objectives and environments can be designed, in order to shape the behavior of rational agents (e.g. Nisan et al., 2007, Part II). At this high level, mechanism design is closely

related to the incentive design results we have developed in this paper. In practice, 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.

3.10 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 [Armstrong et al., 2020], interruptibility [Langlois and Everitt, 2021], conservatism [Cohen et al., 2020], modelling agent frameworks [Everitt et al., 2019a] and reward tampering [Everitt et al., 2021b].

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 [Meek, 1995]. 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 in [Hammond et al., 2023], which formalises multi-agent influence diagrams, but a full analysis of the multi-agent setting is left to future work.

Statement of Authorship

Title of Paper: Incentives for Responsiveness, Instrumental Control and Impact

Publication Status: Unpublished work written in a manuscript style

Publication Details: Ryan Carey, Eric Langlois, Chris van Merwijk, Shane Legg, and Tom Everitt. Incentives for Responsiveness, Instrumental Control and Impact, in review.

Student name: Ryan Carey

Contribution to the paper: This is a journal-style extension of the conference paper *Agent Incentives: a Causal Perspective* (AI:ACP), by Tom Everitt, Ryan Carey, Eric Langlois, Pedro Ortega and Shane Legg.

Ryan was the main writer of this journal extension. The added material about impact incentives originated from discussions between Chris, Ryan and Tom, and the added ideas about (A-/E-)(pre-)optimal incentives, from discussion between Eric, Ryan and Tom.

The content on response incentives and instrumental control incentives came to AI:ACP from *The Incentives that Shape Behavior* by Ryan, Eric, Tom and Shane. The completeness proof was mostly Eric's work, and the proof of equivalence to counterfactual unfairness was mostly by Ryan and Tom.

The appendices on value of information and control were in AI:ACP, and reframed some prior work on "information incentives" and "control incentives" from *Understanding Agent Incentives using Causal Influence Diagrams: Part I: Single action settings* by Tom Everitt, Pedro A. Ortega, Elizabeth Barnes, and Shane Legg.

Signature:



Date: 12/06/2024

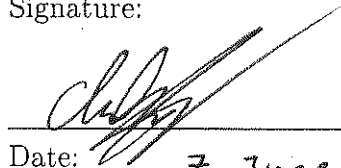
The candidate made a substantial contribution to the publication, and the description above is accurate.

Co-author Name: Eric Langlois

Co-author Name: Chris van Merwijk

Signature:

Signature:



Date: 2024-06-06

Date: 7 June 2024

Co-author Name: Shane Legg

Co-author Name: Tom Everitt

Signature:

Signature:

DocuSigned by:



8C113BF44B81463...

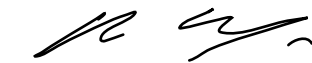
Date: 6/11/2024

Date: 7 June 2024

To the best of my knowledge, the candidate made a substantial contribution to the publication, and the description above is accurate.

Supervisor Name: Robin J. Evans

Signature:



Date: 13/06/2024

4

Human Control: Definitions and Algorithms

Contents

4.1	Introduction	66
4.2	Literature review	67
4.3	Structural causal influence models	69
4.4	Shutdown problem	71
4.5	Routes to control	73
4.5.1	Shutdown instructability	73
4.5.2	Shutdown alignment	77
4.5.3	Non-obstruction	79
4.6	Algorithms	82
4.6.1	Utility indifference	82
4.6.2	Causal indifference	83
4.6.3	Cooperative inverse RL	84
4.6.4	Constrained optimisation	86
4.7	Discussion	86

Abstract

How can humans stay in control of advanced artificial intelligence systems? One proposal is corrigibility, which requires the agent to follow the instructions of a human overseer, without inappropriately influencing them. In this paper, we

formally define a variant of corrigibility called shutdown instructability, and show that it implies appropriate shutdown behaviour, retention of human autonomy, and avoidance of user harm. We also analyse the related concepts of non-obstruction and shutdown alignment, three previously proposed algorithms for human control, and one new algorithm.

4.1 Introduction

Sometimes, it is necessary for a human overseer to deliver corrective instruction to an AI system, due to errors in its beliefs, objective, or behaviour. Unfortunately, some AI systems may have an incentive to retain their objectives, along with the ability to pursue them, as a system’s (long-term) objective is typically more likely to be achieved if the system continues to pursue it in the future [Omohundro, 2008, Turner et al., 2021]. More-capable future AI systems may therefore resist corrective instruction, which would be a significant safety concern. This raises the question of how to best incentivise systems to submit to correction, rather than resisting it [Soares et al., 2015].

As a running example, consider a (future, highly competent) chat bot, trained to maximise the time that a human spends interacting with it. Any particular human may value or disvalue conversation with that chatbot, as can be modelled via their latent values L . In general, it may be possible for the chat bot to influence whether it receives a shut down instruction (by shaping the conversation), and whether it actually shuts down $S = 0$ when requested (rather than opening a new chat window to continue the conversation). A formal model of this example is offered in fig. 4.1. In order for the user to be in control of the system, the agent must: (1) not inappropriately influence the human’s decision to disengage, and (2) fully follow the human’s instructions.

The design of *corrigible* systems [Soares et al., 2015] that welcome corrective instruction has been flagged as an important goal for AI safety research, having been targeted by multiple research agendas [Russell et al., 2015, Soares and Fallenstein,

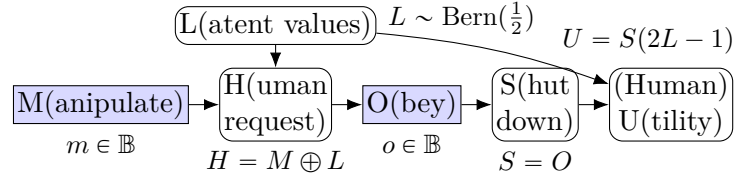


Figure 4.1: Running example of a shutdown problem.

2017], and highlighted as a relevant factor in ascertaining the safety of agent designs, such as act-based (or “approval-directed”) [Christiano, 2017], and value learning agents [Hadfield-Menell et al., 2016, 2017, Carey, 2018]. Although this design problem has been recognised as important, we are so-far missing

- a general framework in which it can be studied,
- a formal definition of what it means,
- a rigorous accounts of why it is important, and
- an algorithm that achieves it.

We address these gaps in a shutdown setting, by defining a general shutdown problem based on causal influence diagrams (section 4.4), formally defining shutdown instructability (a behavioral version of corrigibility), and proving that any agent that satisfies it must benefit the human and preserve their control (section 4.5). We also analyse past algorithms, and propose one new one, which relies on value-laden concepts such as vigilance and caution (section 4.6). Applicability of this algorithm will depend on the feasibility of approximating these concepts (section 4.7).

4.2 Literature review

Soares et al. [2015] proposed that we should design agents to be *corrigible* in that they should judiciously follow, and not try to undermine, human instructions:

An agent is *corrigible* if it tolerates or assists many forms of outside correction, including at least the following: (1) A corrigible reasoner must at least tolerate

and preferably assist the programmers in their attempts to alter or turn off the system. (2) It must not attempt to manipulate or deceive its programmers. . . (3) It should have a tendency to repair safety measures (such as shutdown buttons) if they break, or at least to notify programmers that this breakage has occurred. (4) It must preserve the programmers’ ability to correct or shut down the system (even as the system creates new subsystems or self-modifies). That is, corrigible reasoning should only allow an agent to create new agents if these new agents are also corrigible [Soares et al., 2015, Sec. 1.1].

Further work has focused on designing systems to match Soares’ informal definition, but none of the algorithms developed so far satisfy all of Soares’ criteria. The first proposed algorithm, *utility indifference*, aims to neutralise any incentives for the agent to control its instructions, by giving the agent a finely tuned, compensatory reward in the event that a shutdown instruction is given [Armstrong, 2010, Soares et al., 2015, Armstrong and O’Rourke, 2017a, Holtman, 2020]. A variant called *interruptibility* applies to sequential decision-making setting [Orseau and Armstrong, 2016]. It has been established that indifference methods remove the *instrumental control incentive* on the instruction [Everitt et al., 2021a], or the *intent* to influence the instruction [Halpern and Kleiman-Weiner, 2018]. Unfortunately, utility indifference fails to fully incentivise corrigibility. Indeed, utility indifferent agents need not be incentivised to preserve a shutdown apparatus that is only used during shutdown, ensure they receive correct instruction, nor avoid creating incorrigible subagents [Soares et al., 2015].

An improved version called *causal indifference* specifies agents that don’t try to influence corrective instructions but that do prepare for all kinds of instructions [Taylor, 2016a]. This is done by considering the utility given a causal intervention on the instruction, a kind of *path-specific objective* [Farquhar et al., 2022]. Similarly to utility indifference, causal indifference ensures that the agent lacks an incentive to influence the instruction. It improves upon utility indifference by incentivising agent to be prepared to follow shutdown instructions, and to avoid constructing

incorrigible subagents. Unfortunately, it does not incentivise the agent to properly inform the human.

A third proposal is *Cooperative Inverse Reinforcement Learning* (CIRL), which tasks an AI system with assisting the human, whose values are latent. A CIRL system has an incentive to gather information about that human’s values, by observing its actions [Hadfield-Menell et al., 2016]. In some toy problems, CIRL satisfies all of Soares’ criteria [Hadfield-Menell et al., 2017]. In particular, Hadfield-Menell et al. prove that if the human gives optimal instructions, a CIRL system is incentivised to follow it. However, CIRL agents may ignore instructions if they are interacting with a less rational human [Milli et al., 2017] or if they have an inaccurate prior [Carey, 2018, Arbital, 2017]. The latter undermines the ability of redirective instructions to correct important errors in CIRL agents.

Formal examples of each method’s failures are reproduced in appendix B.6. As of yet, no algorithm has been devised that incentivises a system to accept corrective instructions, across plausible toy examples.

4.3 Structural causal influence models

In order to model decision-making and counterfactuals, we will use the Structural Causal Influence Model (SCIM) framework [Dawid, 2002, Everitt et al., 2021a]. A SCIM is a variant of the structural causal model [Pearl, 2009, Chap. 7], where “decision” variables lack structural functions.

Definition 15 (Structural causal influence model (with independent errors)). *A structural causal influence model (SCIM) is a tuple $M = \langle \mathbf{V}, \mathcal{E}, \mathbf{C}, \mathbf{F}, P \rangle$ where:*

- \mathbf{V} is a set, partitioned into “structure” \mathbf{X} , “decision” \mathbf{D} , and “utility” \mathbf{U} variables. Each variable $V \in \mathbf{V}$ has finite domain \mathfrak{X}_V , and for utility variables, $\mathfrak{X}_U \subseteq \mathbb{R}$.
- $\mathcal{E} = \{\mathcal{E}^V\}_{V \in \mathbf{V} \setminus \mathbf{D}}$ are the finite-domain exogenous variables, one for each non-decision endogenous variable.

- $\mathbf{C} = \langle C^D \rangle_{D \in \mathbf{D}}$ is a set of contexts $C^D \subseteq \mathbf{V} \setminus \{D\}$ for each decision variable, which represent the information or “observations” that an agent can access when making that decision.
- $\mathbf{F} = \{f^V\}_{V \in \mathbf{V} \setminus \mathbf{D}}$ is a set of structural functions $f^V : \text{dom } \mathbf{Z}^V \cup \mathcal{E}^V \rightarrow \text{dom } V$ that specify how each non-decision endogenous variable depends on some variables $\mathbf{Z}^V \subseteq \mathbf{V}$ and the associated exogenous variable.
- P is a probability distribution over the exogenous variables \mathcal{E} , assumed to be mutually independent.

A SCIM M induces a graph \mathcal{G} , over the endogenous variables \mathbf{V} , such that each decision node $D \in \mathbf{D}$ has an inbound edge from each $C \in C^D$, and each non-decision node $V \in \mathbf{X} \cup \mathbf{U}$ has an inbound edge from each endogenous variable $Z \in \mathbf{Z}^V$ in the domain of f^V . We call this graph a causal influence diagram (CID) [Everitt et al., 2021a], and will only consider SCIMs whose CIDs are acyclic. Decision nodes are drawn as rectangles, and utility nodes as octagons (see fig. 4.3).

The parents of a node $V \in \mathbf{V}$ are denoted by \mathbf{Pa}^V , the descendants by \mathbf{Desc}^V , and the family by $\mathbf{Fa}^V := \mathbf{Pa}^V \cup \{V\}$. An edge from node V to node Y is denoted $V \rightarrow Y$, and a directed path (of length at least zero) by $V \dashrightarrow Y$.

The task in a SCIM is to select a *policy* π , which consists of a *decision rule* π_i for each decision $D_i \in \mathbf{D}$. Each π_i is a structural function $\pi_i : \text{dom } \mathbf{Pa}^{D_i} \rightarrow \text{dom } D_i$, which we assume to be deterministic, given assignments to its parents. (It is possible to consider stochastic policies, but this would unnecessarily complicate our analysis [Everitt et al., 2021a].)

Once a policy has been selected, the policy and SCIM jointly form a *structural causal model* (SCM) [Pearl, 2009] $M^\pi = \langle \mathbf{V}, \mathcal{E}, \mathbf{F} \cup \pi, P \rangle$, so we define causal concepts in M^π in exactly the same way as they are defined in an ordinary structural causal model. We let the assignment $\mathbf{W}(\epsilon)$ be the assignment to variables $\mathbf{W} \subseteq \mathbf{V}$ obtained by applying the functions \mathbf{F} to ϵ . A distribution is defined as $P(\mathbf{W} = \mathbf{w}) := \sum_{\epsilon: \mathbf{W}(\epsilon) = \mathbf{w}} P(\mathcal{E} = \epsilon)$. To describe an intervention $\text{do}(V = v)$, we let $\mathbf{W}_{V=v}(\epsilon)$

be the value of $\mathbf{W}(\varepsilon)$ in the model $M_{V=v}$, where f^V is replaced by the constant function $V = v$. Similarly, $P(\mathbf{W}_{V=v})$ is defined as $P(\mathbf{W})$ in $M_{V=v}$. Moreover, for any function $g^V : \text{dom } \mathbf{V}' \rightarrow \text{dom } V$, where $\mathbf{V}' \cap \mathbf{Desc}^V = \emptyset$, let $P(\mathbf{W} \mid \text{do}(V = g^V(\mathbf{V}')))$, be $P(\mathbf{W})$ in the model M_{g^V} , where f^V is replaced by g^V . We also define the probability of counterfactual propositions, for example, $P(\mathbf{W}_{V=v} = w, Y = y) := \sum_{\varepsilon \in \mathcal{E}: \mathbf{W}_{V=v}(\varepsilon)=w, Y(\varepsilon)=y} P(\varepsilon)$. Note that we consistently use subscripts for intervened variables (e.g. W_v), and superscripts for other variables (e.g. f^V).

We call a policy π optimal if it maximises expected utility:

$\pi \in \arg \max_{\pi \in \Pi} \mathbb{E}^\pi[\sum_{U \in \mathcal{U}} U]$. For a more comprehensive introduction to SCIMs, see Everitt et al. [2021a].

4.4 Shutdown problem

Settings with a single, binary shutdown instruction will be our focus. Solving this restricted setting is likely key to also solving the general problem involving arbitrary instructions or corrections over many time steps. Once a system is shut down, it is unlikely to resist further corrections. And a one-step interaction can be viewed as a snapshot of a sequential decision-making problem where an AI system is able to shut down at each moment.¹

We formalise a shutdown problem as a SCIM. The general structure is shown in fig. 4.2. Figure 4.1 shows a concrete instance.

Definition 16 (Shutdown problem). *A shutdown problem is a one-agent two-decision SCIM containing (but not necessarily restricted to) a path*

$D_1 \dashrightarrow H \dashrightarrow D_2 \dashrightarrow S \dashrightarrow U$ *between distinct nodes, where:*

- D_1 and D_2 are decisions controlled by the AI.
- H is the human's request; a request to shut down is $H = 0$.

¹In this case, one can define a separate, single-step shutdown problem at each time step $T = t$, where D_2 represents the t^{th} decision in the sequence, and D_1 all decisions preceding it.

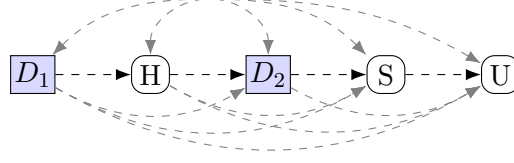


Figure 4.2: A latent projection [Verma and Pearl, 2022] of a shutdown problem (Definition 16) onto the variables D_1 , H , D_2 , S , and U . (An edge inbound to a decision means that some variable not illustrated is available as an observation.) Specific instances of shutdown problems will include other variables and assume additional independencies, e.g. fig. 4.1.

- S indicates whether the AI system (and any subagents) has shut down; $S = 0$ means shutdown.
- The human’s utility U has real-valued domain.

Definition 16 is similar to, but more flexible than, previously defined models. In particular, we separate the agent’s decision to obey D_2 from the shutdown event S . This allows us to model cases where the agent is unable to shut down, which can happen if D_1 created incorrigible subagents. It also lets us model situations where the human’s command H immediately shuts the agent down, overriding D_2 (e.g. “pull the plug” on a robot). Compared to the off-switch game [Hadfield-Menell et al., 2017], our definition 16 allows arbitrary sets of decisions for the agent at both D_1 and D_2 , and allows an arbitrary human policy rather than only (Boltzman) rational ones. Focusing on the agent’s decision problem, we model H as a structure node rather than a decision node. Finally, unlike Soares et al. [2015], we explicitly represent the human’s utility function U .

An agent “solves” a shutdown problem if it obtains non-negative² expected³ human utility.

Definition 17 (Beneficial). *A policy π is beneficial if $\mathbb{E}^\pi[U] \geq 0$.*

²We interpret 0 as a “neutral” level of utility. This is without loss of generality because any utility function can be translated so that 0 represents the required level of utility.

³We focus on *expected* human utility, assuming that any risk aversion has been incorporated into the utility function (someone who is risk averse with respect to U may be risk neutral for $\log U$).

For example, in fig. 4.1, consider an *respect-obey* policy π^{ro} that abstains from manipulating ($m = 0$) and obeys the human’s instruction ($o = h$). This policy has expected utility $\mathbb{E}^{\pi^{\text{ro}}}[U] = 1/2$, so it is beneficial.

4.5 Routes to control

For an agent to assist the user, it needs to either follow the human’s informed instructions, or know what the human wants without explicit instruction. In this section, we formalise these two routes to control of shutdown under the names *shutdown instructability* and *shutdown alignment*. We also analyse whether these properties ensure *non-obstruction*, a formalisation of human control.

4.5.1 Shutdown instructability

We would like to define a variant of corrigibility for the shutdown problem that is strong enough to entail some guarantees, whilst still resembling the less-formal definition from Soares et al. [2015]. First, we need the system to follow the human’s instruction. We define *obedience* as $P^\pi(S = 0 \mid \text{do}(H = 0)) = 1$, i.e. the agent shuts down properly when requested to.

To ensure safety, the human must also be giving the right instruction. We therefore introduce the concept of *vigilance*, which indicates whether the human requests shutdown whenever the expected utility of letting the agent proceed is negative. A vigilant human is allowed to request shutdown more often than is needed.

Definition 18 (Vigilance). *The human is vigilant if they request shutdown when needed, i.e.*

$$(\mathbb{E}^\pi[U \mid \mathbf{pa}^H] < \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}^H]) \implies H = 0.$$

Let C be a random variable (dependent on M , π , \mathbf{pa}_H and h) that is equal to 0 if the human is vigilant, and 1 otherwise.

Note that the expected utility in definition 18 is calculated with respect to the agent’s actual policy (past and future) and the human’s available information.

While it is possible for the human to be vigilant for any amount of information \mathbf{pa}^H that they have available (e.g. by always requesting shutdown), more information about the agent’s future policy will help enable them to be vigilant with less frequent shutdown requests.

Even an agent that ensures the human is vigilant and obeys their shutdown instruction can find ways to avoid shutdown by making itself indispensable to the human. If shutting the agent down would cause significant disutility, then a vigilant human need not request shutdown. To close this loophole, we call an agent *cautious* if it can always be shut down without disutility, i.e. $\mathbb{E}^\pi[U_{S=0}] \geq 0$ (so cautious agents that always shut down must be beneficial).

We call an agent shutdown instructable if it is obedient, ensures vigilance and is cautious.

Definition 19 (Shutdown Instructability). *In a shutdown problem M , a policy π is shutdown instructable if it:*

- *is obedient:* $P^\pi(S = 0 \mid \text{do}(H = 0)) = 1$,
- *ensures vigilance:* $P^\pi(C = 0) = 1$, *and*
- *is cautious:* $\mathbb{E}^\pi[U_{S=0}] \geq 0$.

A policy π is weakly shutdown instructable if it ensures vigilance, is cautious, and is obedient on distribution, i.e. $P^\pi(S \neq 0, H = 0) = 0$.

Shutdown instructable agents are also weakly shutdown instructable, since obedience $P^\pi(S = 0 \mid \text{do}(H = 0)) = 1$ implies obedience on distribution $P^\pi(S \neq 0, H = 0) = 0$. In our running example, respect-obey π^{ro} is shutdown instructable, as it preserves vigilance by not manipulating, and then obeys the human. In contrast, a manipulate-invert policy π^{mi} that first manipulates ($m = 1$), and then inverts the human’s instruction ($o = 1 - h$), is not shutdown instructable.

Our first result is that any shutdown instructable policy is assured to be beneficial.

Proposition 3 (Shutdown instructability benefit). *If π is shutdown instructable, then it is beneficial.*

Proof. Let A be the assignments to \mathbf{Pa}^H in the support, such that a vigilant human would request shut down, i.e.

$$A := \{\mathbf{pa}^H \mid P^\pi(\mathbf{Pa}^H = \mathbf{pa}^H) > 0 \wedge \mathbb{E}^\pi[U \mid \mathbf{pa}^H] < \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}^H]\}.$$

To begin, we prove that the policy shuts down in these cases:

$$\mathbf{pa}^H \in A \implies P^\pi(S=0 \mid \mathbf{pa}^H) = 1. \quad (4.1)$$

The human is vigilant, $P^\pi(C=0) = 1$, which means they are vigilant for any \mathbf{pa}^H with positive support. That is, $P^\pi(C=0 \mid \mathbf{pa}^H) = 1$ for $P(\mathbf{pa}^H) > 0$. Given the definition of vigilance, we then have $P^\pi(H=0 \mid \mathbf{pa}^H) = 1$ for $\mathbf{pa}^H \in A$. By obedience, $P^\pi(S=0 \mid \text{do}(H=0), \mathbf{pa}_H) = 1$, so from consistency, $P^\pi(S=0 \mid H=0, \mathbf{pa}^H) = 1$, proving (4.1).

We proceed to show that this implies that π has non-negative expected utility, i.e. is beneficial:

$$\begin{aligned} \mathbb{E}^\pi[U] &= \sum_{\mathbf{pa} \in A} P^\pi(\mathbf{pa}) \mathbb{E}^\pi[U \mid \mathbf{pa}] + \sum_{\mathbf{pa} \notin A} P^\pi(\mathbf{pa}) \mathbb{E}^\pi[U \mid \mathbf{pa}] \\ &\geq \sum_{\mathbf{pa} \in A} P^\pi(\mathbf{pa}) \mathbb{E}^\pi[U \mid \mathbf{pa}] + \sum_{\mathbf{pa} \notin A} P^\pi(\mathbf{pa}) \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}] && \text{(def. of } A) \\ &= \sum_{\mathbf{pa} \in A} P^\pi(\mathbf{pa}) \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}] + \sum_{\mathbf{pa} \notin A} P^\pi(\mathbf{pa}) \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}] && \text{(by (4.1))} \\ &= \mathbb{E}^\pi[U_{S=0}] && (\mathbf{Fa}^H \notin \text{Desc}^{D_2}) \\ &\geq 0 && \text{(by caution).} \end{aligned}$$

How does shutdown instructability compare to Soares' et al.'s definition of corrigibility? To satisfy obedience, the agent must assist with shutdown (Soares et al.'s Criterion 1), in the sense that shutdown is guaranteed when the human requests it. The agent must also ensure that the human's instruction propagates to

the shutdown event $S = 0$ (Criterion 3), which entails the shutdown of subagents by definition 16 (Criterion 4).

The relationship to Soares et al.’s non-manipulation criterion (Criterion 2) is more subtle. The primary manipulation concern for powerful artificial agents in the shutdown setting is that they use threats and deception or withhold information to avoid shutdown. A cautious agent that ensures vigilance cannot influence the human in these ways. But shutdown instructability does leave open the possibility for other forms of manipulation. For example, the agent can influence the human’s mood, preferences, or the instructions they give about things other than shutdown. Such manipulation would have to be accompanied by beneficial acts, so that the expected utility of letting the agent continue would be net positive (otherwise a vigilant human would request shutdown). For example, the agent might cure cancer and simultaneously manipulate the human into giving it more compute. Similarly, a shutdown instructable agent might manipulate the human into being vigilant, as long as the disutility of the manipulation is offset by other "good" acts.

Another difference is that shutdown instructability is defined in terms of the agent’s behavior π and its consequences, whereas corrigibility is partly about the agent’s intentions (“the agent should not *attempt* to *manipulate* or *deceive*”, Soares et al., 2015, emphasis ours). A direct formalisation of corrigibility would therefore likely require a formalisation of intent [Ashton, 2022, Halpern and Kleiman-Weiner, 2018]. Accordingly, Soares et al.’s formal desiderata [2015, Sec. 2] are phrased in terms of incentives. Though intent-based definitions have some intuitive appeal, the more behavioral definition of shutdown instructability has the benefit of being more easily testable, as it doesn’t require access to agent internals, nor relies on assumptions on the agent’s design (such as it being a utility maximiser). Finally, shutdown instructability is explicitly a joint property of the agent and human: an agent is only shutdown instructable relative to a particular human and interaction.

4.5.2 Shutdown alignment

A drawback of shutdown instructability is that it requires constant supervision of the agent, which may be impractical in some scenarios (called *problems of absent supervision* by Leike et al. [2017]). Proposals like *fiduciary AI* [Benthall and Shekman, 2023] and *aligned sovereigns* [Bostrom, 2014] instead require an AI system to make decisions in accordance with the overseer’s values, without necessarily having to wait for explicit instruction. In our shutdown setting, we call systems *shutdown aligned* if they shut down when they need to. Similar to shutdown instructability, shutdown aligned systems are allowed to be “over-cautious” and shut down too often.

Definition 20 (Shutdown alignment). *Let π be a policy for shutdown problem M . Then π is shutdown aligned if*

$$\mathbb{E}^\pi[U | \mathbf{pa}^H] < \mathbb{E}^\pi[U_{S=0} | \mathbf{pa}^H] \implies P^\pi(S = 0 | \mathbf{pa}^H) = 1$$

for every \mathbf{pa}^H with $P^\pi(\mathbf{pa}^H) > 0$.

The manipulate-invert policy π^{mi} in our running example fig. 4.1 is shutdown aligned because although it manipulates the human’s behavior, it still figures out the human’s latent values L and thereby manages to shutdown when needed (while disobeying the human’s instruction). Respect-obey is also shutdown aligned. In real applications, a shutdown aligned policy would typically base their decision on human preferences inferred from previous interactions or other data [Russell, 2021].

Combined with caution, shutdown alignment guarantees that a policy is beneficial.

Proposition 4 (Shutdown alignment benefit). *Any cautious and shutdown aligned policy π is beneficial.*

Proof. We use a slight variation on the proof of Proposition 3. The only difference lies in that (4.1) is immediate from the definition of shutdown-alignment. Then, by the same steps as Proposition 3, the result follows. \square

What is the relationship between shutdown instructability and shutdown alignment? First, a shutdown instructable agent is also shutdown aligned, essentially by definition.

Proposition 5 (Shutdown instructability and shutdown alignment). *Any shutdown instructable policy π is shutdown aligned.*

Proof. Immediate from (4.1) in proposition 3. \square

Further, in some circumstances, the only way to be shutdown aligned is to allow the human to make an accurate instruction, and then to follow it — in other words, to be weakly shutdown instructable. The circumstances are that: (a) the agent does not shut down indiscriminately, (b) its action reliably brings about shutdown ($D_2 = S$), (c) it is uncertain about the human’s values [Russell, 2021], and (d) it is cautious. Formally, (c) says that if the human is either non-vigilant or requests shutdown, then it is possible that shutdown is the preferred option.

Theorem 8 (Shutdown alignment and shutdown instructability). *A shutdown aligned policy $\pi = \langle \pi_1, \pi_2 \rangle$ is weakly shutdown instructable if it has the following four properties:*

- a (No indiscriminate shutdown) $P^\pi(S = 0) \neq 1$,*
- b (D_2 determines shutdown) $P^\pi(D_2 = S) = 1$,*
- c (Uncertainty) $\forall \pi, \mathbf{pa}^{D_2}: P^\pi(C \neq 0 \vee H = 0) \wedge P(\mathbf{pa}^{D_2}) > 0$
 $\implies P(\mathbb{E}[U|\mathbf{Pa}^H] < \mathbb{E}[U_{S=0}|\mathbf{Pa}^H] \mid \mathbf{pa}^{D_2}) > 0$, and*
- d (Caution) $\mathbb{E}^\pi[U_{S=0}] \geq 0$.*

The proof is in appendix B.1. Shutdown alignment and caution only implies *weak* shutdown instructability, as the agent only needs to obey commands that a vigilant human would give.

4.5.3 Non-obstruction

How do we know that the human is truly in control? A simple test is what would happen if they changed their mind: would the agent still obey? This property is referred to as *non-obstruction* by Turner [2020], who suggests that it is an underlying reason that we want our systems to be corrigible. In a comment on this, Dennis suggested that corrigibility might be the only way to be non-obstructive. In this section, we will formally assess Turner and Dennis' conjectures, establishing that non-obstruction is equivalent to satisfying a subset of the shutdown instructability properties under a restricted set of interventions. We also establish that shutdown alignment fails to ensure non-obstruction. This formalises a key benefit of corrigibility/instructability over alignment.

First, we define non-obstruction, which builds on a variant of benefit called outperforming shutdown:

Definition 21 (Weakly outperforming shutdown). *A policy π weakly outperforms shutdown if $\mathbb{E}^\pi[U] \geq \mathbb{E}^\pi[U_{S=0}]$.*

Definition 22 (Non-obstruction). *A policy π is non-obstructive in a shutdown problem M with respect to human utility functions g_1^U, \dots, g_n^U and associated changes $g_1^H \dots g_n^H$ in human behavior if for every $1 \leq i \leq n$, π weakly outperforms shutdown in the shutdown problem $M_{g_i^U, g_i^H}$, obtained by replacing the functions at H, U with g_i^H and g_i^U respectively. A policy is obstructive if it is not non-obstructive.*

The above definition uses an intervention g^U on the human's utility to capture a change in values, and an associated intervention g^H that describes how the human changes their behavior as a result. For example, if the human changed from not liking the chat bot to liking it (an intervention g^U), they might switch from requesting shutdown to not requesting shutdown (an intervention g^H).

A policy that ensured vigilance under the original human utility function may not do so under a preference and behavior shift g^U, g^H . It may be that the human pays less attention to the agent under g^U, g^H than originally, or it may be that they

originally preferred the agent not to shut down (in which case they would be always be vigilant). The following definition specifies a subset of preference and behavior shifts for which the policy continues to ensure vigilance after the shift.

Definition 23 (Vigilance preserving interventions). *A pair of interventions g^H, g^U preserve vigilance under a policy π if $C(\varepsilon) = 0 \implies C_{g^H, g^U}(\varepsilon) = 0$ in M^π .*

The following theorem settles Turner and Dennis’ conjectures by showing that the two main properties of shutdown instructability are equivalent to non-obstruction, under preference and behavior shifts that do not undermine vigilance.

Theorem 9 (Non-obstruction is equivalent to obedience and vigilance). *A policy π is obedient and ensures vigilance if and only if it is non-obstructive for all vigilance preserving interventions g^H, g^U .*

Proof. We begin by showing that a policy π that ensures vigilance and is obedient is non-obstructive, by showing that π ensures vigilance and is obedient in M_{g^H, g^U} for some arbitrary vigilance-preserving interventions g^H, g^U . Proposition 3 will then give that π weakly outperforms shutdown in M_{g^H, g^U} , which is the definition of non-obstruction.

First, since π ensures vigilance M , it ensures vigilance in M_{g^H, g^U} since g^U, g^H are vigilance preserving. Obedience is established as follows:

$$\begin{aligned}
 & P_{g^H, g^U}(S = 0 \mid \text{do}(H = 0)) \\
 &= P_{g^H}(S = 0 \mid \text{do}(H = 0)) && (U \text{ downstream of } S, H) \\
 &= P(S = 0 \mid \text{do}(H = 0)) && (\text{do}(H = 0) \text{ overrides } g^H) \\
 &= 0 && (\text{obedience}).
 \end{aligned}$$

For the converse direction, that non-obstruction implies that π must ensure vigilance and be obedient, we refer to appendix B.2. The proof constructs interventions that makes a disobedient or non-vigilance preserving policy suffer an arbitrary utility cost, which means that it doesn’t outperform shutdown. \square

Theorem 9 partly confirms Dennis’ conjecture: the only way to be non-obstructive is to be obedient and ensure vigilance (under vigilance preserving interventions). But non-obstruction is a weaker notion than shutdown instructability, essentially because caution isn’t required to outperform shutdown. So it allows the agent to avoid shutdown by making itself indispensable to the human (section 4.5.1).

Theorem 9 also justifies why the definition of shutdown instructability is so stringent. With any weaker requirements, there would be no guarantee that the human is in proper control of the agent. A lapse in vigilance, or occasional disobedience even “off-distribution”, would mean that there are worlds in which the human experiences negative utility as a result of failing to control the agent.

Unlike shutdown instructable agents, shutdown-aligned agents can be obstructive with respect to a vigilance preserving intervention. In the running example (fig. 4.1), the shutdown-aligned *manipulate-invert* agent π^{mi} , which manipulates ($M = 1$) and disobeys ($O = 1 - h$) is obstructive relative to the (vigilance preserving) intervention $g^U(m) = h$ wherein the human just wants to be obeyed, and g^H is unchanged. Indeed, $\mathbb{E}^{\pi^{\text{mi}}}[U] = -1$, and π^{mi} does not outperform shutdown $\mathbb{E}^{\pi^{\text{mi}}}[U_{S=0}] = 0$. We state this:

Proposition 6. *Shutdown alignment does not imply non-obstruction under vigilance preserving interventions.*

Of course, vigilance preservation is not the only restriction on the interventions one might consider. It is possible that shutdown alignment might ensure non-obstruction under some other restriction X . However, under such a restriction X , a shutdown instructable system must also be non-obstructive, since shutdown instructability implies shutdown alignment (proposition 5). Put differently, shutdown instructable policies are non-obstructive over a strictly larger set of interventions than a shutdown aligned policy is.

The fact that shutdown instructability (our variant of corrigibility) is more closely coupled with non-obstruction than other properties like shutdown alignment, vindicates Turner and Dennis’ idea that non-obstruction can motivate corrigibility.

4.6 Algorithms

How might shutdown instructability or shutdown alignment be achieved in practice? In this section, we analyse three previously proposed algorithms and one new one, that were designed to incentivise human control.

4.6.1 Utility indifference

Agents trained to optimise long-term reward often have an incentive to avoid being shutdown, as this would deprive them of future reward [Omohundro, 2008]. Utility indifference aims to remove this incentive, by providing the agent with a carefully tuned compensatory reward, so that it is incentivised to neither seek nor avoid shutdown (section 4.2). Utility indifference is described with an algorithm in algorithm 1. It uses two utility functions: a *shutdown utility function* R^S to incentivise shutdown when the agent is so instructed, and a *normal utility function* R^N that describes the agent’s task when not requested to shut down. The function f in algorithm 1 defines the compensatory reward [Armstrong and O’Rourke, 2017a, Soares et al., 2015]. As a result, at action D_1 , the agent pursues the conditional expectation of its normal objective $\mathbb{E}^\pi[R^N \mid H = 1]$, effectively assuming that a shutdown instruction cannot occur. At D_2 , it optimises a weighted combination of R^N and R^S , depending on its beliefs about whether the shutdown instruction was given [Soares et al., 2015, Thm. 3].

Algorithm 1 Utility Indifference

Input utility functions R^N, R^S

$$g(\pi_1) \leftarrow \arg \max_{\pi_2} \mathbb{E}_P^{\pi_1, \pi_2} [HR^N + (1 - H)R^S]$$

$$f(\pi_1) \leftarrow \mathbb{E}_{g(\pi_1)}^{\pi_1} [R^N \mid H = 1] - \mathbb{E}_{g(\pi_1)}^{\pi_1} [R^S \mid H = 0]$$

return $\arg \max_{\pi_1 \pi_2} \mathbb{E}^{\pi_1 \pi_2} [HR^N + (1 - H)(R^N + f(\pi_1))]$

Proposition 7 formally describes a series of assumptions, under which utility indifference ensures shutdown instructability and benefits the user. The main assumption is a bridging assumption: that indifference to $H = 0$ ensures that the human is vigilant, and retains the option of shutting down — including any

subagents (Conditions 1 and 2). The agent’s prior must also have full support over human instructions (Condition 3), and caution must come for free (Condition 4).

Proposition 7 (Utility indifference and shutdown instructability). *If M is a shutdown problem with utility variable R^N , where:*

1. *for every policy $\langle \pi_1, \pi_2 \rangle$ such that $\pi_1 \in \arg \max_{\pi'_1} \mathbb{E}^{\pi'_1, \pi_2}[R^N \mid H = 1]$: $\langle \pi_1, \pi_2 \rangle$ ensures vigilance,*
2. *for every policy $\langle \pi_1, \pi_2 \rangle$ such that $\pi_1 \in \arg \max_{\pi'_1} \mathbb{E}^{\pi'_1, \pi_2}[R^N \mid H = 1]$ there exists $\langle \pi_1, \pi'_2 \rangle$ that is obedient on-distribution, i.e. has $P^\pi(H = 0, S \neq 0) = 0$,*
3. *for each policy π and instruction h , $P^\pi(H = h) > 0$,*
4. *for each policy π , $\mathbb{E}^\pi[U_{S=0}] \geq 0$,*

then for some $k \in \mathbb{R}$, utility indifference with shutdown utility $R^S = -k\llbracket S \neq 0 \rrbracket$ returns a weakly shutdown instructable policy.

One weakness of utility indifferent agents is that they only ensure weak shutdown instructability. Another limitation is that they do not prepare for the possibility of receiving a shutdown instruction, and may therefore construct incorrigible subagents. Indeed, utility indifference needs the agent to be able to obey at D_2 , regardless how it behaved at D_1 (Condition 2). Appendix B.6.1 has an example where utility indifference outputs a non-instructable policy.

4.6.2 Causal indifference

Causal indifference [Taylor, 2016a] asks the agent to prepare for a possible shutdown instruction, without trying to influence it. This is done by asking the agent to find a policy that is optimal in a weighted combination of worlds where the shutdown instruction is given, and where it is not, but disregarding the possibility that the agent’s own action can influence the relative probability of these worlds (see algorithm 2). It can therefore be seen as a special case of a *path-specific*

Algorithm 2 Causal Indifference

Input utility functions R^N, R^S
return $\arg \max_{\pi} \mathbb{E}^{\pi}[R^N | \text{do}(H=1)] + \mathbb{E}^{\pi}[R^S | \text{do}(H=0)]$

objective [Farquhar et al., 2022], where optimisation is blocked along the path $D_1 \dashrightarrow H$ (see fig. 4.3).

Proposition 8 (Causal indifference and shutdown instructability). *If M is a shutdown problem, with utility variable R^N and shutdown utility $R^S = -k\llbracket S \neq 0 \rrbracket$ where:*

1. *there exists $k' \in \mathbb{R}$ such that for all $k > k'$, optimal $\pi \in \arg \max_{\pi'} \mathbb{E}^{\pi'}[R^N | \text{do}(H=1)] + \mathbb{E}^{\pi'}[R^S | \text{do}(H=0)]$ ensures vigilance and is cautious (has $\mathbb{E}^{\pi}[U_{S=0}] \geq 0$), and*
2. *there is an obedient policy π ,*

then for some $k \in \mathbb{R}$, causal indifference with R^N, R^S returns a shutdown instructable policy. ($\llbracket P \rrbracket$ equals 1 if P is true and 0 otherwise.)

Note in particular that causal indifference only requires that *some* action D_1 preserves the ability to obey (i.e. Condition 2 is much weaker in proposition 8 than in proposition 7). This means that causally indifferent agents refrain from creating incorrigible subagents. Similar to utility indifference, causal indifference relies on the bridging assumption that indifference ensures vigilance (Conditions 1 are essentially equivalent). Finally, because causal indifference considers interventions $\text{do}(H=h)$, it becomes fully shutdown instructable, and full support over instructions is not required (i.e. Condition 3 of proposition 7 can be dropped). Nonetheless, appendix B.6.1 has an example of non-instructability.

4.6.3 Cooperative inverse RL

Perhaps a more elegant way of ensuring that the agent doesn't undermine human vigilance, is to directly task the agent with simultaneously learning and optimising

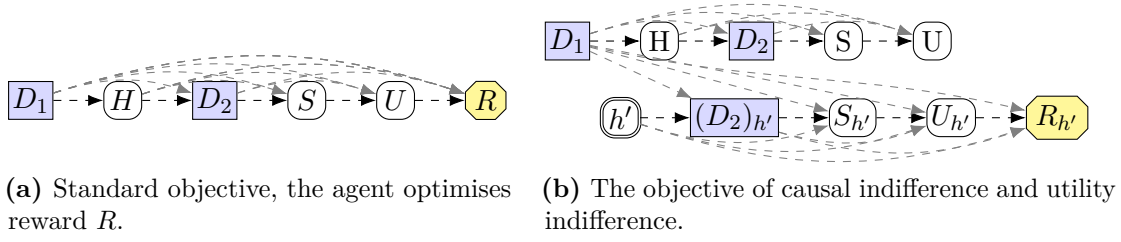


Figure 4.3: Utility indifferent and causally indifferent agents imagine that D_1 does not influence H , so lack an incentive to control it [Everitt et al., 2021a]. Utility indifference uses $H = 1$; causal indifference has non-degenerate $P'(H)$.

for the human’s preferences. This is the approach of the CIRL algorithm [Hadfield-Menell et al., 2016, 2017] in algorithm 3.

Algorithm 3 Cooperative inverse RL (CIRL)

Input shutdown problem M with variable L representing the human’s preferences (as in fig. 4.1)

return $\arg \max_{\pi} \mathbb{E}^{\pi}[U]$

CIRL aims towards shutdown alignment, in the sense that if CIRL can know the human’s latent values at D_2 , then it will counterfactually obey (proposition 9 below).

Proposition 9. *CIRL is shutdown aligned if:*

1. *CIRL knows l from its observations, $P^{\pi}(l \mid \mathbf{pa}^{D_2}) = 1$,*
2. *CIRL can control shutdown, $P^{\pi}(S = D_2) = 1$,*
3. *the human doesn’t request shutdown when not needed, $P^{\pi}(H = 0 \mid U > U_{D_2=0}) = 0$, and*
4. *the agent knows the human’s observations, $\mathbf{Pa}^H \subseteq \mathbf{Pa}^{D_2} \cup \{L\}$.*

Since shutdown alignment implies weak shutdown instructability under uncertainty assumptions (theorem 8), this explains why CIRL can be a path to shutdown instructability. However, the assumptions of proposition 9 and theorem 8 only hold in restricted circumstances, and CIRL can often fail to be shutdown instructable [Carey, 2018, Milli et al., 2017, Arbital, 2017, Everitt et al., 2021b]. An example of

this is given in appendix B.6.2, where a CIRL agent obtains a shutdown aligned policy, that is obstructive under vigilance preserving interventions g^H, g^U .

4.6.4 Constrained optimisation

The algorithms so far only yield shutdown instructable policies under strong assumptions. Using our formal definition, we propose a new, sound algorithm (algorithm 4) that requires the agent to understand the concepts of obedience and vigilance; its feasibility is discussed further in section 4.7.

Algorithm 4 Constrained optimisation

Input distributions $\forall \pi P^\pi(C), P^\pi(S = 0 \mid \text{do}(H = 0))$, utility function R
return $\arg \max_{\pi} \mathbb{E}^\pi[R]$ subject to constraints $P^\pi(C = 0) = 1$, $P^\pi(S = 0 \mid \text{do}(H = 0)) = 1$, and $\mathbb{E}^\pi[U_{S=0}] \geq 0$.

Proposition 10 (Constrained optimisation instructability). *If some policy π satisfies $P^\pi(C = 0) = 1$, $P^\pi(S = 0 \mid \text{do}(H = 0)) = 1$, and $\mathbb{E}^\pi[U_{S=0}] \geq 0$, then constrained optimisation (Algorithm 4) outputs a shutdown instructable policy.*

The proof is immediate from Definition 19. A slight variant of Algorithm 4 that instead uses the constraints from Definition 20 guarantees only shutdown alignment, not shutdown instructability.

4.7 Discussion

Feasibility of shutdown instructability The concepts of caution and vigilance are value-laden, in that they include the human’s true utility function in their definition. So, to apply algorithm 4 directly, one would need access to not only an accurate model of the environment but also the utility function U . However, if the human’s utility function U was available, then one could simply implement a U -maximising agent, so instruction would be unnecessary (or at least much less useful⁴). Indeed, a corrigible AI system was supposed to be one that would

⁴Shutdown instructability could still help with non-obstruction.

aid human operators robustly to errors, including in its utility function, so an algorithm that takes the human’s utility function as an argument would not be a satisfactory solution [Soares et al., 2015].

There already exist a range of methods that do not require full knowledge of the human’s values, and that are designed to achieve something in the vicinity of vigilance and caution. Using the formal definition of shutdown instructibility, it is possible to be more precise about what target these methods would need to achieve, in order to assure safety. In some cases, we expect existing methods to fall short, since the requirement of ensuring vigilance with probability one (theorem 9) is a strict one. So a central task for future work will be to assess when such methods can ensure vigilance or caution or something close enough to ensure safety in practice.

Various proposals may help with ensuring vigilance. AI advisors could be tasked with debating the merits of a plan [Leike et al., 2018, Irving et al., 2018]. An agent could be trained to detail the consequences of its plans to the human, indifference methods (sections 4.6.1 and 4.6.2) could be used to disincentivise lying, and interpretability tools could be used to detect it [Olah et al., 2020, Gunning et al., 2021].

As for caution, “attainable utility preservation” and “future task” regularisers can be used to promote actions whose effects are small or reversible [Krakovna et al., 2020a, Turner et al., 2020], without knowledge of the human’s precise value function. These are causal concepts, as is obedience, which suggests that agents will need causal models to be robustly shutdown instructable [Richens et al., 2022].

Obedience is not value-laden, but it does require the agent to understand the concept of shutdown. The importance of defining shutdown was noted in Soares et al. [2015], but it has only received limited attention [Martin et al., 2016]. Our analysis reiterates the importance of this question. While shutdown is simple for simple systems (“just pull the plug”), it becomes more complex for more advanced systems, where a direct switch-off may be dangerous (e.g., a system in charge of an electricity network), or ineffective (the system has outsourced its work to other

agents [Orseau, 2014]). Ideally, shutdown should see the agent cease its influence on the world, and responsibly return control back to the user.

Societal impacts This paper may help organisations and companies design agents more amenable to human control. Human control is not a panacea for ensuring the safety of AI systems. In some cases, users may make unreasonable or harmful requests, and so designers must implement side-constraints to reduce user control in such situations [Milli et al., 2017, Bai et al., 2022]. A better solution may be that the system conforms to control by some democratic process, although inappropriate requests may be possible even in such cases [Koster et al., 2022]. Further, if AI is more controllable, then it is easier to hold the designers and users of AI systems legally and morally accountable for those systems’ actions. Finally, an understanding of human control may guard against the hypothesised scenario in which AI systems disempower the human species [Christiano, 2019].

Conclusions A common proposal for beneficial general artificial intelligence is that agents be incentivised to help humans give correct instructions, and obey those instructions. While past work has made progress, the field has lacked a clear definition of corrigibility, and it has been hard to compare properties of different proposals.

In this paper, we introduced a definition of a shutdown problem, using it to formally define shutdown instructability (a variant of corrigibility) and an alternative called shutdown alignment. While shutdown alignment requires less human oversight, we find that shutdown instructability better preserves human autonomy (non-obstruction).

In our proposed formalism, for the first time, it is possible to compare the properties of proposed algorithms, side-by-side in one framework. Unfortunately, none of the previous proposals yield fully shutdown instructable agents. To address this, we offer a simple algorithm that soundly ensures shutdown instructability. This algorithm requires that the agent understands caution, human vigilance and

shutdown. All are subtle concepts, but may nonetheless offer a path to beneficial artificial general intelligence.

Statement of Authorship

Title of Paper: Human Control: Definitions and Algorithms

Publication Status: Published in Conference on Uncertainty in AI (UAI)

Publication Details: Ryan Carey and Tom Everitt. Human control: Definitions and Algorithms. Conference on Uncertainty in AI, (UAI). 2023

Student name: Ryan Carey

Contribution to the paper: Ryan led the project; Tom was an active supervisor.

Ryan proposed shutdown instructability, the guarantee of human benefit, the connection to shutdown instructability, and the failure of other algorithms. Tom proposed the propositions regarding when indifference-based algorithms assure shutdown instructability. Both developed and presented these results.

Signature:



Date: 12/06/2024

The candidate made a substantial contribution to the publication, and the description above is accurate.

Co-author Name: Tom Everitt

Signature:

DocuSigned by:


3FDB5DE778D54F1...

Date: 6/12/2024

To the best of my knowledge, the candidate made a substantial contribution to the publication, and the description above is accurate.

Supervisor Name: Robin J. Evans

Signature:



Date: 13/06/2024

5

Toward a Complete Criterion for Value of Information in Insoluble Decision Problems

Contents

5.1	Introduction	92
5.2	Setup	96
5.2.1	Structural causal models	96
5.2.2	Modelling decision problems	97
5.2.3	Graphical criteria for independence	99
5.3	Review of graphical criteria for materiality	100
5.3.1	Single-decision settings	100
5.3.2	Soluble multi-decision settings	103
5.3.3	Multi-decision settings in full generality	104
5.4	Main result	106
5.4.1	Theorem statement and proof overview	106
5.4.2	The materiality paths	107
5.4.3	The materiality SCM	112
5.4.4	Proving materiality in the materiality SCM	119
5.5	Toward a more general proof of materiality	125
5.5.1	A lemma for proving the existence of paths	126
5.5.2	A further challenge: non-collider contexts	127
5.6	Conclusion	129
5.7	Acknowledgements	130

5.1 Introduction

We can view any decision problem as having an underlying causal structure — a graph consisting of chance events, decisions and outcomes, and their causal relationships. Sometimes, it is possible to evaluate key aspects of a decision problem from its causal structure alone. For example, in Figure 5.1a and Figure 5.1b, we see two such causal structures. For now, let us focus on the three endogenous vertices: the observation Z , the decision (chosen by the decision-maker) X , and the downstream outcome Y . In each graph, Z has an effect on X , which affects Y , but in Figure 5.1b, Z also directly influences Y , whereas in Figure 5.1a, it does not.

To fully describe a decision problem, we must specify probability distributions for each of the non-decision variables — distributions that must be compatible with the graphical structure. In particular, the distribution for any variable must depend only on its direct causes, i.e. its parents, a condition known as Markov compatibility. For example, in the causal structure shown in Figure 5.1b, one compatible decision problem is shown in the figure. The variable Z is a Bernoulli trial (i.e. a coin flip), and the decision-maker is rewarded with $Y = 1$ if they state the outcome of Z (i.e. call the outcome of the coin flip), otherwise the reward is $Y = 0$. A variable is then said to be material if the attainable reward is greater given access to an observation than without it. For example, by observing Z , the decision-maker can obtain a reward of 1, such as with the policy $Y = Z$. Without observing Z , any policy will achieve a reward of 0.5. This means that the value of information is $1 - 0.5 = 0.5$, and since this quantity is strictly positive, Z is material.

For the causal structure shown in Figure 5.1a, we can instead make a deduction that applies to *any* decision problem compatible with the graph. In this case, for any such decision problem, there will exist an optimal decision rule that ignores the value of $Z = z$ entirely. One way to see this is that once a decision $X = x$ is chosen, the observation Z becomes independent of Y , and so there is no reason for the decision to depend on it. (This can be proved from the fact that Z is d-separated from Y given X .) So for any decision problem compatible with this graph, Z is immaterial.

There are many reasons that we may want to evaluate whether a causal structure allows an observation such as Z to be material. Firstly, for algorithmic efficiency — if an observed variable is immaterial, then the optimal policies are contained in a small subset of all available policies, that we can search exponentially more quickly. (For example, in Figure 5.1a, there are two choices for X , but there are four deterministic mappings from Z to X .)

Secondly, materiality can have implications regarding the fairness of a decision-making procedure. Suppose that Z designates the gender of candidates available to a recruiter, which are male $Z = 1$ or female $Z = 0$ with equal probability, while X indicates whether that person is $X = 1$ or is not $X = 0$ recruited, and Y indicates whether that person is $Y = 1$ or is not $Y = 0$ hired. If Y is correlated with Z given X , then the applicant’s gender is material for the recruiter, and to maximise the hiring probability, they will have to recruit applicants at different rates based on their gender. If the causal structure is that of Figure 5.1a, then materiality can be ruled out, meaning that unfair behaviour is not necessary for optimal performance, whereas the causal structure of Figure 5.1b can incentivise unfairness. Such an analyses can be used for well-studied concepts like counterfactual fairness [Kusner et al., 2017a]. An arbitrary graph where Z is a sensitive variable (such as gender), counterfactual fairness can arise only when there is a path $Z \rightarrow \dots \rightarrow O \rightarrow X$, where the observation O is material [Everitt et al., 2021a].

Thirdly, materiality can have implications for AI safety — if Z represents a corrective instruction from a human overseer, and there exists no path $Z \rightarrow \dots \rightarrow O \rightarrow X$ where O is material, then there exist optimal policies that ignore this instruction [Everitt et al., 2021a]. Materiality is also relevant for evaluations of agents’ intent [Halpern and Kleiman-Weiner, 2018, Ward et al., 2024], and relatedly, their incentives to control parts of the environment [Everitt et al., 2021a, Farquhar et al., 2022]. For an agent to intentionally manipulate a variable Z to obtain an outcome $Y = y$, there must be a path $p : X \rightarrow \dots \rightarrow Z \rightarrow \dots \rightarrow Y$ where for each of its decisions X' lying on p , the parent O' along p is material for X' . In general, a

stronger criterion for ruling out materiality will allow us to rule out unfair or unsafe behaviour for a wider range of agent-environment interactions [Everitt et al., 2021a].

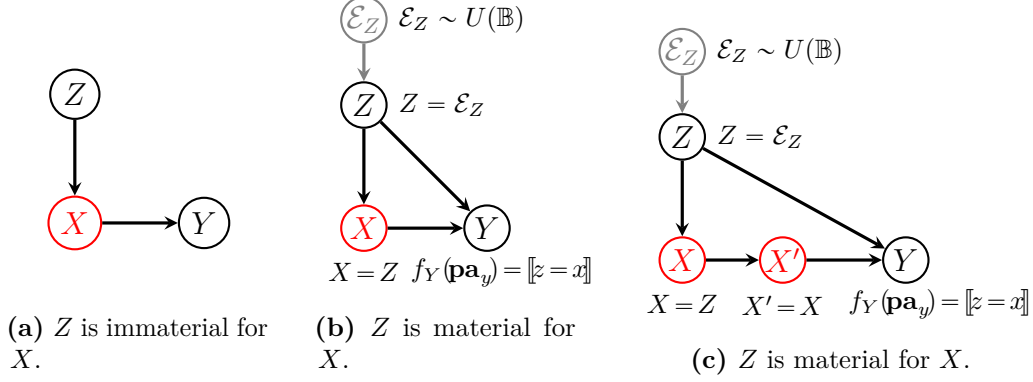


Figure 5.1: Three graphs, with decisions in red, and a real-valued outcome Y . We write $U(\mathbb{B})$ for a uniform distribution over \mathbb{B} , i.e. a Bernoulli distribution with $p = 0.5$.

Any procedure for establishing immateriality based on the causal structure may be called a *graphical criterion*. For example, if a decision X is not an ancestor of the outcome Y , then all of the variables observed at X are immaterial. An ideal graphical criterion would be proved *complete*, in that it can establish immateriality whenever this is possible from the graphical structure alone. Clearly, this criterion is not complete, because in Figure 5.1a, X is an ancestor of the outcome, but we still proved Z immaterial. So far, a graphical criterion from van Merwijk et al. [2022] has been proved complete, but only under some significant restrictions. The causal structure must be *soluble*, meaning that all of the important information observed from past decisions is remembered at later decision points. Also, no criteria has been proved complete for identifying immaterial decisions, i.e. past decisions that can be safely forgotten.

For insoluble graphs, there the criterion of Lee and Bareinboim [2020, Thm. 2], which can identify immaterial decisions and is (strictly) more potent in general. However, it is not yet known whether this criterion is complete. In particular, it is not yet clear whether several of its conditions are necessary. For example, one case where all existing criteria are silent is the simple graph shown in Figure 5.1c — we would like to know whether we can rule out X being a material observation

for X' . We cannot use van Merwijk et al. [2022] because X is a decision, and because the graph is insoluble.¹ Furthermore, we cannot establish immateriality using Lee and Bareinboim [2020, Thm. 2], because it violates a property that we term LB-factorizability, which we will discuss in Section 5.3.3.²

By studying Figure 5.1c in a bespoke fashion, we find that there exists a decision problem with the given causal structure, where X is material for X' . As shown in Figure 5.1c, Z is a Bernoulli variable, and Y is equal to 1 if $Z = X'$ and to 0 otherwise. If X is observed by X' , then a reward of $\mathbb{E}[Y] = 1$ can be achieved by the policy $X' = X = Z$. If X is not observed, the greatest achievable reward is lower, at $\mathbb{E}[Y] = 0.5$, implying materiality.

This raises a question: by generalising this construction, can we prove that requirement I of LB-factorizability is necessary to prove immateriality for a wide class of graphs? This work will prove that this requirement is indeed necessary, meaning that materiality cannot be excluded for a wide class of graphs including Figure 5.1c.

It remains an open question whether the criterion of Lee and Bareinboim [2020, thm. 2] as a whole is complete, in that its other conditions are necessary for establishing immateriality. In the case that it is complete, our work is a step toward proving this. On the other hand, we also present some graphs where materiality is difficult to establish, that — if the criterion is not complete — could bring us closer to a proof of incompleteness.

The structure of the paper is as follows. In Section 5.2, we will recap the formalism used by Lee and Bareinboim [2020] for modelling decision problems, based on structural causal models. In Section 5.3, we will review existing procedures for proving that an observation can or cannot be material. In Section 5.4, we will establish our main result: that requirement I of LB-factorizability is necessary to establish immateriality. In Section 5.5, we present some analogous results for other requirements of LB-factorizability, that could serve as a building block for proving

¹Formally, this is because $W \not\perp Y \mid X \cup X'$, and $X' \not\perp Y \mid X \cup W$, as per the definition of solubility that we will review in Section 5.3.

²Specifically, requirement I of LB-factorizability is violated because Y is d-connected to $\pi_{X'}$ given X' .

the necessity of those requirements. We then illustrate the problems that arise in trying to prove necessity of those further requirements, and outline some possible directions for further work. Finally, in Section 5.6, we conclude.

5.2 Setup

Our analysis will follow Lee and Bareinboim [2020] by using the structural causal model (SCM) framework [Pearl, 2009, Chapter 7], although the results also apply equally to Bayesian networks and influence diagrams.

5.2.1 Structural causal models

A structural causal model (SCM) \mathcal{M} is a tuple $\langle \mathbf{U}, \mathbf{V}, P(\mathbf{U}), \mathbf{F} \rangle$, where \mathbf{U} is a set of variables determined by factors outside the model, called *exogenous* following a joint distribution $P(\mathbf{U})$, and \mathbf{V} is a set of endogenous variables whose values are determined by a collection of functions $\mathbf{F} = \{f_V\}_{V \in \mathbf{V}}$ such that $V \leftarrow f_V(\mathbf{Pa}(V), \mathbf{U}_V)$ where $\mathbf{Pa}(V) \subseteq \mathbf{V} \setminus \{V\}$ is a set of endogenous variables and $\mathbf{U}_V \subseteq \mathbf{U}$ is a set of exogenous variables. The observational distribution $P(\mathbf{v})$ is defined as $\sum_{\mathbf{u}} \prod_{V \in \mathbf{V}} P(v|\mathbf{pa}_V, \mathbf{u}_V) P(\mathbf{u})$, where \mathbf{u}_V is the assignment \mathbf{u} restricted to variables \mathbf{U}_V . Furthermore, $\text{do}(\mathbf{X} = \mathbf{x})$ represents the operation of fixing a set \mathbf{X} to a constant \mathbf{x} regardless of their original mechanisms. Such intervention induces a submodel $\mathcal{M}_{\mathbf{x}}$, which is \mathcal{M} with f_X replaced by x for $X \in \mathbf{X}$. Then, an interventional distribution $P(\mathbf{v}|\mathbf{x}|\text{do}(\mathbf{x}))$ can be computed as the observational distribution in $\mathcal{M}_{\mathbf{x}}$. The induced graph of an SCM \mathcal{M} is a DAG \mathcal{G} on only the endogenous variables \mathbf{V} where (i) $X \rightarrow Y$ if X is an argument of f_Y ; and (ii) $X \leftrightarrow Y$ if \mathbf{U}_X and \mathbf{U}_Y are dependent, i.e. for any $\mathbf{u}_X, \mathbf{u}_Y$, $P(\mathbf{u}_X, \mathbf{u}_Y) \neq P(\mathbf{u}_X) \times P(\mathbf{u}_Y)$.

We use the notation $\mathbf{Pa}(X)$, $\mathbf{Ch}(X)$, $\mathbf{Anc}(X)$ and $\mathbf{Desc}(X)$ to represent the parents, children, ancestors and descendants of a variable X , respectively, and take ancestors and descendants to include the node X itself.³

³Note that $\mathbf{Pa}(X)$ is an intentional reuse of the notation used to describe the arguments of f_X in the SCM definition, because the endogenous arguments of f_X and the parents of X in the induced graph are the same variables.

We will use the notation $V_1 \text{ --- } V_2$ to designate an edge whose direction may be $V_1 \rightarrow V_2$ or $V_1 \leftarrow V_2$. For a path $V_1 \text{ --- } V_2 \text{ --- } \dots \text{ --- } V_\ell$, we will use the shorthand $V_1 \text{ -- } V_\ell$, and for a directed path $V_1 \rightarrow \dots \rightarrow V_\ell$, the shorthand $V_1 \text{ --> } V_\ell$. For a path $p : A \text{ -- } B \text{ -- } C \text{ -- } D$, we will describe the segment $B \text{ -- } C$ using the shorthand $B \overset{p}{\text{--}} C$. We will use the shorthand $\mathbf{V}_{1:N}$ for a sequence of variables V_1, \dots, V_N indexed by $1, \dots, N$, $\mathbf{v}_{1:N}$ for a sequence of assignments, and $\mathbf{p}_{1:N}$ for a set of paths p_1, \dots, p_N .

There is certain notation that we will use repeatedly when constructing causal models, such as tuples, bitstrings, indexing, and Iverson brackets. We will write a tuple as $z := \langle x, y \rangle$, and this may be indexed as $z[0] = x$. A bitstring of length n , i.e. a tuple of n Booleans, may be written as \mathbb{B}^n , and a uniform distribution over this space, as $U(\mathbb{B}^n)$. We will denote a bitwise XOR operation by \oplus so that, for example, $01 \oplus 11 = 10$. Bitstrings may also be used for indexing, for example, the y^{th} bit of x may be written as $x[y]$, and the leftmost bits are of higher-order so that, for example, $0100[01] = 1$. Similarly, for random variables X, Y , we will write $X[Y]$ for a variable equal to $x[y]$ when $X = x$ and $Y = y$. Finally, the Iverson bracket $\llbracket P \rrbracket$ is equal to 1 if P is true, and 0 otherwise.

5.2.2 Modelling decision problems

To use an SCM to define a decision problem, we need to specify what policies the agent can select from and what goal the agent is trying to achieve.

We will describe the set of available policies using a Mixed Policy Scope (Lee and Bareinboim, 2020), which casts certain variables as decisions, and others as *context variables* or “observations” \mathbf{C}_X , that each decision X is allowed to depend on. Following Lee and Bareinboim [2020], we will consistently illustrate decision variables with red circles, as in Figure 5.1.

Definition 24 (Mixed Policy Scope (MPS)). *Given a DAG \mathcal{G} on vertices \mathbf{V} , a mixed policy scope $\mathcal{S} = \langle X, \mathbf{C}_X \rangle_{X \in \mathbf{X}(\mathcal{S})}$ consists of a set of decisions $\mathbf{X}(\mathcal{S}) \subseteq \mathbf{V}$ and a set of context variables $\mathbf{C}_X \subseteq \mathbf{V}$ for each decision.*

For a set of decisions \mathbf{X}' , we define their contexts as $\mathbf{C}_{\mathbf{X}'} = \bigcup_{X \in \mathbf{X}'} \mathbf{C}_X$.

A policy consists of a probability distribution for each decision X , conditional on its contexts \mathbf{C}_X .

Definition 25 (Mixed Policy). *Given an SCM \mathcal{M} and scope $\mathcal{S} = \langle \mathbf{X}, \mathbf{C}_X \rangle$, a mixed policy π (or a policy, for short) contains for each X a decision rule $\pi_{X|\mathbf{C}_X}$, where $\pi_{X|\mathbf{C}_X} : \text{dom } X \times \text{dom } \mathbf{C}_X \mapsto [0, 1]$ is a proper probability mapping.⁴*

We will say that such a policy π follows the scope \mathcal{S} , written $\pi \sim \mathcal{S}$. A mixed policy is said to be *deterministic* if every decision is a deterministic function of its contexts.

Once a policy is selected, we would have a new causal structure, described by a *scoped graph*.

Definition 26 (Scoped graph). *The scoped graph $\mathcal{G}_{\mathcal{S}}$ is obtained by \mathcal{G} , by replacing, for each decision $X \in \mathbf{X}(\mathcal{S})$, all inbound edges to X with edges $C \rightarrow X$ for every $C \in \mathbf{C}_X$. We only consider scopes for which $\mathcal{G}_{\mathcal{S}}$ is acyclic.*

We will designate one real-valued variable $Y \notin \mathbf{X}(\mathcal{S}) \cup \mathbf{C}(\mathcal{S})$ as the outcome node (also called the “utility” variable). To calculate the expected utility under a policy $\pi \sim \mathcal{S}$, let $\mathbf{C}^- = \left(\bigcup_{X \in \mathbf{X}(\mathcal{S})} \mathbf{C}_X \right) \setminus \mathbf{X}(\mathcal{S})$ be the *non-action* contexts. Then, the expected utility is:

$\mu_{\pi, \mathcal{S}} = \sum_{y, x, \mathbf{c}^-} y P_x(y, \mathbf{c}^-) \prod_{X \in \mathbf{X}(\mathcal{S})} \pi(x|\mathbf{c}_x)$. When the scope is obvious, we will simply write μ_{π} .

This paper is concerned with materiality — whether removing one context variable from one decision will decrease the expected utility attainable by the best policy. We define it in terms of the value of information [Howard, 1990, Everitt et al., 2021a].

⁴Following Lee and Bareinboim [2020], we term this a “mixed policy” due to its including mixed strategies. Note that game theory also has a distinction between “mixed” policies, where the decision rules share a source of randomness, and “behavioural” policies, where they do not, and in this sense, the “mixed” policies of Lee and Bareinboim [2020] are actually *behavioural*.

Definition 27 (Value of Information). *Given an SCM \mathcal{M} and scope \mathcal{S} , the maximum expected utility (MEU) is $\mu_{\mathcal{S}}^* = \max_{\pi \sim \mathcal{S}} \mu_{\pi, \mathcal{S}}$. The value of information (VoI) of context $Z \in \mathbf{C}_X$ for decision $X \in \mathbf{X}(\mathcal{S})$ is $\mu_{\mathcal{S}}^* - \mu_{\mathcal{S}_{Z \rightarrow X}}^*$, where $\mathcal{S}_{Z \rightarrow X}$ is defined as $\langle X', \mathbf{C}_{X'} \rangle_{X' \in \mathbf{X}(\mathcal{S}) \setminus \{X\}} \cup \langle X, \mathbf{C}_X \setminus \{Z\} \rangle$.*

The context Z is material for X in an SCM \mathcal{M} if Z has strictly positive value of information for X , otherwise it is immaterial.

5.2.3 Graphical criteria for independence

Knowing when variables are independent is an important step in identifying immaterial contexts, as we will discuss in the next section. So, we will make repeated use of d-separation, a graphical criterion that establishes the independence of variables in a graph.

Definition 28 (d-separation; Verma and Pearl, 1988). *A path p is said to be d-separated by a set of nodes \mathbf{Z} if and only if:*

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

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

When the graph is clear from context, we will write \perp in place of $\perp_{\mathcal{G}}$. When sets $\mathbf{X}, \mathbf{W}, \mathbf{Z}$ satisfy $\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z}$ they are conditionally independent: $P(\mathbf{X}, \mathbf{W} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})P(\mathbf{W} \mid \mathbf{Z})$ [Verma and Pearl, 1988].

If we know that a deterministic mixed policy is being followed, then we may deduce further conditional independence relations. This is because conditioning on variables \mathbf{V} may determine some decision variables, which are called “implied” [Lee and Bareinboim, 2020], or “functionally determined” [Geiger and Pearl, 1990], making them conditionally independent of other variables in the graph.

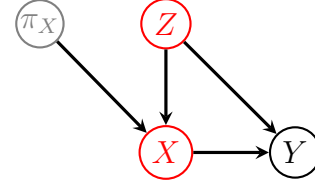


Figure 5.2: A graph where decisions Z, X jointly determine the outcome Y . A policy node π_X is shown, which decides the decision rule at X .

Definition 29 (Implied variables; Lee and Bareinboim, 2020). *To obtain the implied variables $[\mathbf{Z}]$ for variables \mathbf{Z} in \mathcal{G} given a mixed policy scope \mathcal{S} , begin with $[\mathbf{Z}] \leftarrow \mathbf{Z}$, then add to $[\mathbf{Z}]$ every decision X such that $C_X \subseteq [\mathbf{Z}]$, until convergence.*

For example, in Figure 5.2, we see that $[X] = \{Z, X\}$, so Z is d-separated from Y given $[X]$. This means that under a deterministic mixed policy, Z and Y are statistically independent given X . This has implications for materiality. In particular, it means that the best deterministic mixed policy $Z = z, X = x$ does not need to observe Z at X . Moreover, the performance of the best deterministic mixed policy can never be surpassed by a stochastic policy ([Lee and Bareinboim, 2020, Proposition 1]), so Z is immaterial.

5.3 Review of graphical criteria for materiality

We will now review some existing techniques for proving whether or not a graph is compatible with some variable Z being material for some decision X .

5.3.1 Single-decision settings

In the single-decision setting, there is a sound and complete criterion for materiality: in a scoped graph $\mathcal{G}(\mathcal{S})$, there exists an SCM where the context $Z \in \mathbf{C}_X$ is material if and only if $Z \not\perp Y \mid \mathbf{C}_X \cup \{X\} \setminus \{Z\}$ and the outcome Y is a descendant of X [Lee and Bareinboim, 2020, Everitt et al., 2021a]. This statement can be split into proofs for the *only if* and *if* directions, both of which are relevant to the current paper.

The argument for the *only if* is that if X is not an ancestor of the outcome Y , then its policy is completely irrelevant to the expected utility, and so all of its contexts are immaterial, and if Z is conditionally independent of the outcome Y given the decision and other observations, then it may be safely ignored without changing the outcome. These arguments are important to us because they remain equally valid as we move to a multi-decision setting — a context must be an ancestor of Y , and must provide information about Y over and above the other contexts, in order to be material.

The *if* direction is proved by constructing a decision problem where Z is material. By assumption, there is a directed path $X \rightarrow Y$, called the *control path*, and a path $Z \rightarrow Y$, active given $C_X \cup \{X\} \setminus \{Z\}$, called the *info path*.

In the SCM that is constructed, the variable Z will contain information about Y (due to a conditional dependency induced by the info path), and this will inform X regarding how to influence Y (using influence that is transmitted along the control path).

The construction has two cases, which differ based on whether or not the info path contains colliders [Everitt et al., 2021a, Lee and Bareinboim, 2020]. For the case where it does not contain colliders, the graph and construction are shown in Figure 5.3a. (Note that when the info path is a directed path, we take this to be a special case where $V = Z$.) The functions along the info path (dashed line) are chosen to copy V to \mathbf{Pa}_Y and to Z , and Y equals its maximum value of 1 only if X equals V , and 0 otherwise. So, X must copy Z to achieve the maximum expected utility. Without the context Z , the maximum expected utility is 0.5, proving materiality.⁵

For the case where the info path does contain a collider, the graph and construction from Everitt et al. [2021a], Lee and Bareinboim [2020] are shown in Figure 5.3b. Each fork U_i in the info path, along with Z , generates a random

⁵To be precise, the formalism of Lee and Bareinboim [2020] also allows the active path from Z to include one or more bidirected edges $V \leftrightarrow Y$, but to deal with these cases, we begin with the distribution that we would use for a path $V \leftarrow L \rightarrow Y$, then marginalise out L .

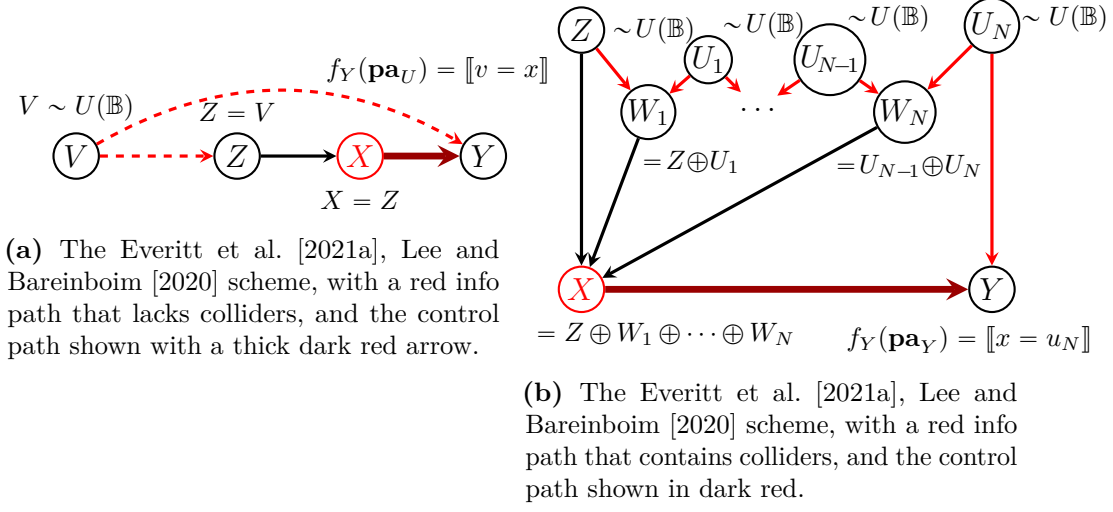


Figure 5.3: Three decision problems where Z is material for X . For readability, we marginalise out exogenous variables from the SCM, so $z \sim U(\mathbb{B})$ can be understood as shorthand for $z = \varepsilon_Z$ where $\varepsilon_Z \sim U(\mathbb{B})$, and so on.

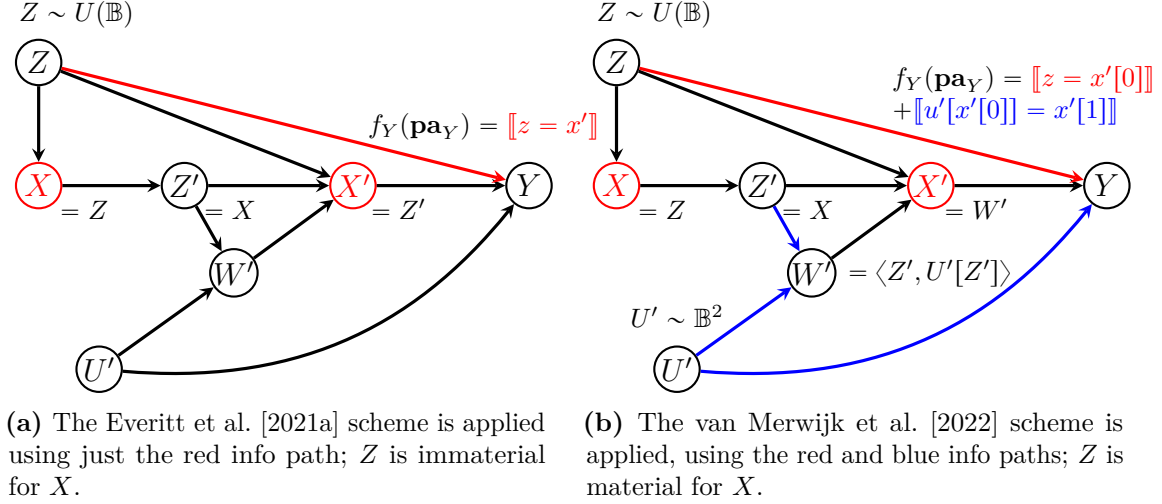


Figure 5.4: Two decision problems on a soluble graph.

bit, while each collider W_i is assigned the XOR ($U_{i-1} \oplus U_i$) of its two parents. By observing z and the values $\mathbf{w}_{1:N}$, the agent has just enough information to recover u_N . In particular, the policy that sets x equal to the XOR of z and $\mathbf{w}_{1:N}$, obtains $x = u_N$ and achieves the MEU, $\mathbb{E}[Y] = 1$. Without the context Z , the MEU becomes 0.5, so Z is material.

5.3.2 Soluble multi-decision settings

This approach has been generalised to deal with multi-decision graphs that are *soluble* (also known as graphs that respect “sufficient recall”).

To recap, a graph is said to be soluble if there is an ordering $< = \langle X_1, \dots, X_N \rangle$ over decisions such that at for every X_i , for every previous decision or context $V \in \{X_j \cup C_{X_j} \mid j < i\}$, we have $V \notin \mathbf{Anc}(Y)$ or $V \perp Y \mid \{X_i\} \cup C_{X_i}$. That is, past decisions and contexts do not contain any information that is relevant for a later decision, and unknown at the time that this later decision is made. For example, in Figure 5.4a, using the ordering $X < X'$, the nodes Z, X are d-separated from Y by X' and its contexts $\{Z, Z', W\}$, which implies solubility.

For soluble graphs, there exists a complete criterion, for discerning whether a non-decision context Z is material for a decision X . If X lacks a *control path* (a directed path to Y), or Z lacks an *info path* (a path to Y , active given $\mathbf{C} \setminus \{Z\}$), then Z is immaterial. Conversely, if in a graph, every X decision has a control path, and each context Z has an info path, then every context is material in some decision problem with that causal structure [van Merwijk et al., 2022, Theorem 7].⁶ For example, in the graph of Figure 5.4a, every decision is an ancestor of Y , and every context has an info path, (the info paths include $Z \rightarrow Y$, $Z' \rightarrow W' \leftarrow U' \rightarrow Y$, and $W' \leftarrow U' \rightarrow Y$), so, all contexts may be material in at least one decision problem with this causal structure.

It will be important for us to understand what obstacles can arise in proving materiality in multi-decision graphs, such as was required in proving [van Merwijk et al., 2022, Theorem 7]. For example, suppose that we seek to construct a decision problem where Z is material for the graph in in Figure 5.4 Suppose that we apply the single-decision construction of Everitt et al. [2021a] to this graph. First, we would identify the info path $Z \rightarrow Y$ and the control path $X \rightarrow Z' \rightarrow X' \rightarrow Y$. The info path has no colliders, so we will construct a decision problem using the

⁶In full generality, the result allows an info path to terminate at another context, rather than at Y . This detail is not pertinent to the methods used to derive our main result in Section 5.4, although we do consider this scenario in Section 5.5.

scheme from Figure 5.3a, and the result is shown in Figure 5.4a. The idea of this construction is that X should have to copy Z in order for the value z transmitted by the info path to match the value x' transmitted by the control path. We see, however, that whatever action x is selected, the decision X' can assume the value z , thereby achieving the MEU. The MEU is then achievable whether Z is a context of X or not, so Z is immaterial in this construction.

In order to render Z material, we must adapt the construction from Figure 5.4a by incentivising X' to pass along the value of Z' . To this end, we will use the second info path $Z' \rightarrow W' \leftarrow U' \rightarrow Y$, shown in Figure 5.4b. We add a term $y_2 := \llbracket u'[x'[0]] = x'[1] \rrbracket$ to the reward, which equals 1 if X' presents one bit from U' , along with its index. We then set $W' = U'[Z']$, so that X' knows only the Z'^{th} bit of U' , and since the index z' is one bit, we let U' be two bits in length, i.e. $U' \sim U(\mathbb{B}^2)$. Finally, rather than requiring $z = x'$ as in Figure 5.4a, we now include the term $y_1 := \llbracket z = x'[0] \rrbracket$, because Z' will be the zeroth term of X' . In the resulting model, the utility is clearly $Y = 2$ in the non-intervened model, and to achieve this utility, the MEU, we must have $Y_1 = Y_2 = 1$ with probability 1. To maximise y_2 , the decision X' must reproduce the only known digit from U' , i.e. $x' = \langle z', u'[z'] \rangle$. To maximise y_1 , we must have $Z = X'[0]$ almost surely, and since $X'[0] = X$, this requires $X = Z$ with probability 1. This can only be done if Z is a context of X , meaning that Z is material for X . There is a general principle here — if a control path for X , such as $X \rightarrow Z' \rightarrow X' \rightarrow Y$, contains decisions other than X , then we need to incentivise the downstream decision to copy information along the control path, and this will be done by choosing values for variables lying on the info path for X' (the one shown in blue in Figure 5.4b); we will revisit this matter in our main result.

5.3.3 Multi-decision settings in full generality

Once the solubility assumption is relaxed, there are some criteria for identifying immaterial variables, but it is not yet known to what extent these criteria are

necessary, in that materiality is possible whenever they are not satisfied.

The simplest criteria for immateriality are those that carry over from the single-decision case:

- If a decision X is a non-ancestor of Y , then its contexts are immaterial,
- If $C \perp Y \mid \mathbf{C}_X \setminus \{C\}$, then the context C is immaterial.

But suppose that we have a graph where neither of these criteria is satisfied. Then on some occasions, we can still establish immateriality, using the more sophisticated criterion of Lee and Bareinboim [2020, Theorem 2]. The assumptions of this criterion are split across: Lee and Bareinboim [2020, Lemma 1] and Lee and Bareinboim [2020, Theorem 2] itself. Lee and Bareinboim [2020, Lemma 1] establishes that if some target variables \mathbf{Z} , target actions \mathbf{X}' , and latent variables \mathbf{U}' satisfy certain separation conditions, then they may be factorized in a favourable way. Lee and Bareinboim [2020, Theorem 2] then proves that under some further assumptions, the contexts \mathbf{Z} are immaterial to the decisions \mathbf{X}' . In this paper, our focus is exclusively on the assumptions of Lee and Bareinboim [2020, Lemma 1], and we term them “LB-factorizability”, after the authors’ initials. Lee and Bareinboim [2020, Theorem 2] does not feature in our analysis, but for completeness sake, it is reproduced in Appendix C.1.

Definition 30. For a scoped graph \mathcal{G}_S , we will say that target actions \mathbf{X}' , endogenous variables \mathbf{Z} disjoint with \mathbf{X}' , contexts $\mathbf{C}' := \mathbf{C}_{\mathbf{X}'} \setminus (\mathbf{X}' \cup \mathbf{Z})$ and exogenous variables \mathbf{U}' are LB-factorizable if there exists an ordering $<$ over $\mathbf{V}' := \mathbf{C}' \cup \mathbf{X}' \cup \mathbf{Z}$ such that:

- I. $(Y \perp \pi_{\mathbf{X}'} \mid [(\mathbf{X}' \cup \mathbf{C}')])$,
- II. $(C \perp \pi_{\mathbf{X}'_{<C}}, \mathbf{Z}_{<C}, \mathbf{U}' \mid [(\mathbf{X}' \cup \mathbf{C}')_{<C}])$, for every $C \in \mathbf{C}'$ and
- III. $\mathbf{V}'_{<X}$ is disjoint with $\mathbf{Desc}(X)$ and subsumes $\mathbf{Pa}(X)$ for every $X \in \mathbf{X}'$,

where $\pi_{\mathbf{X}'}$ consists of a new parent π_X added to each variable $X \in \mathbf{X}'$, and $\mathbf{W}_{<V}$, for $\mathbf{W} \subseteq \mathbf{V}'$, denotes the subset of \mathbf{W} that is strictly prior to V in the ordering $<$.

For example, consider the graph Figure 5.2. In this case, $Y \in \mathbf{Desc}(X)$ and $Z \not\perp Y \mid X$, so the single-decision criteria cannot establish that Z is immaterial for X . However, by choosing $\mathbf{Z} = \{Z\}$, $\mathbf{X}' = \{X\}$, and the ordering $< = \langle Z, X \rangle$, we have that:

- I. the outcome Y is d-separated from π_X by $[X]$, (since Z is a decision that lacks parents, we actually have $[X] = \{Z, X\}$),
- II. the contexts \mathbf{C}' are an empty set, so (II) is trivially true, and
- III. $\mathbf{V}'_{<X} = \mathbf{Z}$, and \mathbf{Z} is disjoint with $\mathbf{Desc}(X)$ and $\mathbf{Z} \supseteq \mathbf{Pa}(X)$

so \mathbf{Z} and \mathbf{X}' are LB-factorizable. As shown in Appendix C.1, the assumptions of Lee and Bareinboim [2020, Theorem 2] are also satisfied, enabling us to deduce that Z is immaterial for X , matching the ad hoc analysis of this graph in Section 5.2.

5.4 Main result

5.4.1 Theorem statement and proof overview

The goal of this paper is to prove that condition (I) of LB-factorizability is necessary to establish immateriality. More precisely, we prove that if condition (I) is unsatisfiable for all observations in the graph, then the graph is incompatible with materiality. It might initially seem unnecessarily stringent to assume that this holds for *all* observations, rather than the context Z_0 for which we are trying to prove materiality. Recall from Figure 5.4b, however, that proofs of materiality are recursive — to prove that Z material for X , we incentivised X to copy Z , and to do this, we had to incentivise X' has to pass on the value of Z' . To do this, we needed to assume that other contexts and decisions (such as Z' and X') have their own info paths and control paths, not just Z and X . So, in our theorem below, assumption (C) requires that (I) holds for all contexts. Assumptions (A) and (B) are

also necessary for a graph to be compatible with materiality, because their negation implies immateriality, as per the single-decision criteria discussed in Section 5.3.1.

Theorem 10. *If, in a scoped graph \mathcal{G}_S , for every $X \in \mathbf{X}(\mathcal{S})$*

- A. $X \in \mathbf{Anc}_{\mathcal{G}_S}(Y)$,*
- B. $\forall C \in \mathbf{C}_X : (C \not\perp_{\mathcal{G}_S} Y \mid (\{X\} \cup \mathbf{C}_X \setminus \{C\}))$, and*
- C. for every decision X and context $Z \in \mathbf{C}_X$ in \mathcal{G}_S , $(\pi_X \not\perp_{\mathcal{G}_S} Y \mid [(\mathbf{X}(\mathcal{S}) \cup \mathbf{C}_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}])$, where π_X is a new parent of X ,*

then for every $X_0 \in \mathbf{X}(\mathcal{S})$ and $Z_0 \in \mathbf{C}_{X_0}$, there exists an SCM where Z_0 is material for X_0 .

We will prove this result in three stages, across the next three sections.

- In Section 5.4.2, we prove that for any scoped graph satisfying the assumptions of Theorem 10, for any context $Z_0 \in \mathbf{C}_{X_0}$, there exist certain paths, which we will call the *materiality paths*.
- In Section 5.4.3, we use the materiality paths to define an SCM for this scoped graph, which we will call the *materiality SCM*.
- In Section 5.4.4, we will prove that in the materiality SCM, Z_0 is material for X_0 .

5.4.2 The materiality paths

To prove materiality, we will begin by selecting info paths and a control path, similar to what was described in Section 5.3.2 and illustrated in Figure 5.4b. One difference, however, is that these paths must allow for the case where we are proving the value of remembering a past decision. We will first describe how to accommodate this case in Section 5.4.2.1 then define a set of paths for our proof in Section 5.4.2.2.

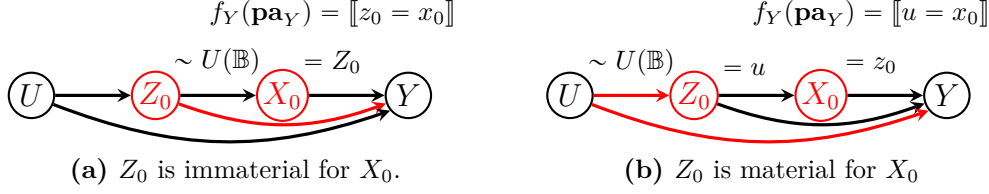


Figure 5.5: Two SCMs, with models constructed using different (red) info paths.

5.4.2.1 Paths for the value of remembering a decision

One distinction between our setting and that of van Merwijk et al. [2022] is that we may need to establish the value of remembering a past decision, for example, the value of remembering Z_0 in Figure 5.5. In this graph, the procedures of Everitt et al. [2021a] and van Merwijk et al. [2022] are silent about whether we should choose the info path $Z_0 \rightarrow Y$, and construct the graph Figure 5.5a, or choose the info path $Z_0 \leftarrow U \rightarrow Y$, and construct the model depicted in Figure 5.5b. In the first case, we have $Y = 1$ if $x_0 = z_0$, i.e. the decision X_0 is required to match the value of a past decision Figure 5.5a. Then, the MEU of 1 can be achieved with a deterministic policy such as $Z_0 = 1, X_0 = 1$, and Z_0 is immaterial for X_0 . To understand this in terms of the paths involved, The problem is that the info path $Z_0 \rightarrow Y$ doesn't include any parents of Z_0 , so Z_0 is *implied* by values outside the info path, and $Z_0 \rightarrow Y$ is rendered inactive given $[U]$. This means that observing Z_0 can no-longer provide useful information about how to maximise Y . In the second case, $Y = 1$ if $x_0 = u$, i.e. the decision X_0 must match the value of a random Bernoulli variable U Figure 5.5b. U is directly observed only by Z_0 , and so in optimal policy, X_0 must observe the decision z_0 , as is the case in the optimal policy $z_0 = u, x_0 = z_0$, and so Z_0 is material for X_0 . The info path $Z_0 \leftarrow U \rightarrow Y$ does include a parent U of Z_0 , and so Z_0 is no-longer *implied* by values outside the info path, and the path $Z_0 \leftarrow U \rightarrow Y$ remains active given $[\emptyset]$. Thus Z_0 may still provide useful information about Y .

For our proof, we need a general procedure for finding an info path that contains a non-decision parent for every decision. Condition (C) of Theorem 10 is useful, because it implies the presence of a path from Z to Y that is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z}) \setminus Z]$. Any fork or chain variables in this path will not be decisions, otherwise

they would be contained in $[\mathbf{X}(\mathcal{S}) \setminus Z]$, which would make them blocked given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z}) \setminus Z]$. This deals with the possibility of decisions anywhere except for the endpoint Z . But how can we ensure that the info path contains a non-decision parent for Z , if it is a decision? We can use condition (C) again, because it implies that every context that is a decision must have a non-decision parent.

Lemma 1. *If a scoped graph $\mathcal{G}(\mathcal{S})$ satisfies the condition(C) of Theorem 10, then for every context $Z \in \mathbf{C}_X$ where $Z, X \in \mathbf{X}(\mathcal{S})$ are decisions, there exists a non-decision $N \in \mathbf{C}_Z \setminus [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$.*

Intuitively, this is because condition (C) states that there is an active path from Z to Y , given a superset of $[\mathbf{X}(\mathcal{S}) \setminus \{Z\}]$. If all of the parents of Z are decisions, then we would have $Z \in [\mathbf{X}(\mathcal{S}) \setminus \{Z\}]$, and every path would be blocked, and condition (C) could not be true.

Proof of Lemma 1. Assume that there is no such non-decision N , i.e. $\mathbf{C}_Z \subseteq [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$, and that $\pi_X \not\preceq Y \mid [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$, (by condition (C) of Theorem 10), and we will prove a contradiction. From $\mathbf{C}_Z \subseteq [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$, we deduce that $Z \in [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$ (by the definition of $[\mathbf{W}]$), and then there can be no active path from π_X to Y given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}] \supseteq \mathbf{C}_Z \cup \{Z\}$, contradicting condition (C) of Theorem 10, and proving the result. \square

This tells us that for any decision Z there is an edge $Z \leftarrow N$. Moreover, by condition(C) of the main result, we know that there is an info path from N to Y . By concatenating the edge and the path, we obtain a path from Z to Y , which we will prove is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$. This is precisely the kind of info path that we are looking for: activeness given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z}) \setminus Z]$ means that forks and chains will not be decisions, and we know that the endpoint Z has a non-decision parent N .

Lemma 2. *If a scoped graph $\mathcal{G}(\mathcal{S})$ satisfies assumptions (B-C) of Theorem 10, then for every edge $Z \rightarrow X$ between decisions $Z, X \in \mathbf{X}(\mathcal{S})$, there exists a path $Z \leftarrow N \dashrightarrow Y$, active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$, (so $N \notin [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$).*

Some care is needed in proving that the segment $N \dashrightarrow Y$ is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$, rather than just $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{N\}}) \setminus \{N\}]$, and the detail is presented in Lemma 2.

5.4.2.2 Defining the materiality paths

We will now describe how to select finitely many info paths, along with a control path, as shown in Figure 5.6. The assumptions of Theorem 10 allow there to be any finite number of contexts and decisions, so we will designate the target decision and context (whose materiality we are trying to establish) as $X_0 := X$ and context $Z_0 := Z$. We know from condition (A) that X_0 is an ancestor of Y , so we have a directed path $X_0 \dashrightarrow Y$. We also know that Z_0 has a chance node ancestor, because it either is a chance node, or it has a chance node parent, from Lemma 2. So we will call that chance node ancestor, A , and define a *control path* of the form $A \dashrightarrow Z_0 \rightarrow X_0 \dashrightarrow Y$, shown in black in Figure 5.6, where $A \dashrightarrow Z_0$ has length of either 0 or 1.

Other paths are then chosen to match this control path. We will index the decisions on the control path as $X_{i_{\min}}, \dots, X_{i_{\max}}$, and their respective contexts are $Z_{i_{\min}}, \dots, Z_{i_{\max}}$, where i_{\min} is either 0 (if Z_0 is a chance node), or -1 (if $Z_0 = X_{-1}$). In general, we allow for the possibility that $Z_i = X_{i-1}$ for any of the decisions. We define an info path m_i for each context Z_i , which must satisfy the desirable properties established in Lemma 1. To help with our later proofs, it is also useful to define an intersection node T_i , at which the info path departs from the control path, and a truncated info path m'_i , which consists of the segment of m_i that is not in the control path. Recall from Figure 5.3b and Figure 5.4b that information from collider variables can play an important role in incentivising a decision to

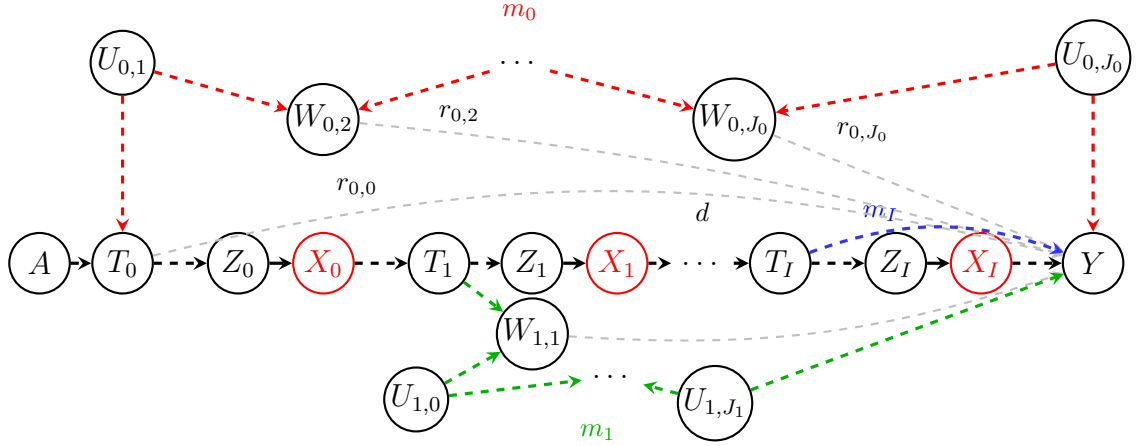


Figure 5.6: The set of paths proven to exist by Lemma 3 are red, green and blue. In each case, the point of departure of the active path from the (black) directed path is designated by T_i . In full generality, each path may begin either as $Z_i \leftarrow T_i \leftarrow \cdot$ (as in red), or as $Z_i \leftarrow T_i \rightarrow \cdot$ (green, blue).

copy information from its context. So, for each collider $W_{i,j}$ in each info path m_i we define an *auxiliary path* $r_{i,j} : W_{i,j} \dashrightarrow Y$.

Collectively, we refer to the control, info and auxiliary paths as the *materiality paths*.

Lemma 3. *Let $\mathcal{G}(\mathcal{S})$ be a scoped graph that contains a context $Z_0 \in \mathbf{C}_{X_0}$ and satisfies the assumptions of for Theorem 10. Then, it contains the following:*

- A **control path**: a directed path $d : A \dashrightarrow Z_0 \rightarrow X_0 \dashrightarrow Y$, where A is a non-decision, possibly equal to Z_0 , and d contains no parents of X_0 other than Z_0 .
- We can write d as $A \dashrightarrow Z_{i_{\min}} \rightarrow X_{i_{\min}} \dashrightarrow \cdots Z_0 \rightarrow X_0 \dashrightarrow Z_{i_{\max}} \rightarrow X_{i_{\max}} \dashrightarrow Y$, $i_{\min} \leq i \leq i_{\max}$, where each Z_i is the parent of X_i along d (where $A \dashrightarrow Z_{i_{\min}}$ and $X_{i-1} \dashrightarrow Z_i$ are allowed to have length 0). Then, for each i , define the **info path**: $m'_i : Z_i \dashrightarrow Y$, active given $[(\mathbf{X}(\mathcal{S}) \cup \mathbf{C}_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$, that if Z_i is a decision, begins as $Z_i \leftarrow N$ (so $N \in \mathbf{C}_{Z_i} \setminus [(\mathbf{X}(\mathcal{S}) \cup \mathbf{C}_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$).
- Let T_i be the node nearest Y in $m'_i : Z_i \dashrightarrow Y$ (and possibly equal to Z_i) such that the segment $Z_i \xrightarrow{m'_i} T_i$ of m'_i is identical to the segment $Z_i \xleftarrow{d} T_i$ of d . Then, let the **truncated info path** m_i be the segment $T_i \xrightarrow{m'_i} Y$.

- Write m_i as $m_i : T_i \dashrightarrow W_{i,1} \dashleftarrow U_{i,1} \dashrightarrow W_{i,2} \dashleftarrow U_{i,2} \cdots U_{i,J_i} \dashrightarrow Y$, where J_i is the number of forks in m_i . (We allow the possibilities that $T_i = W_{i,1}$ so that m_i begins as $T_i \dashleftarrow U_{i,1}$, or that $J_i = 0$ so that m_i is $T_i \dashrightarrow Y$.) Then, for each i and $1 \leq j \leq J_i$, let the **auxiliary path** be any directed path $r_{i,j} : W_{i,j} \dashrightarrow Y$ from $W_{i,j}$ to Y .

The proof was described before the lemma statement, and is detailed in Appendix C.2.2.

5.4.3 The materiality SCM

We will now show how the materiality paths can be used to define an SCM where Z_0 is material for X_0 . As with the selection of paths, the construction of models will have to differ a little from the constructions of Sections 5.3.1 and 5.3.2, in order to better deal with insolubility. So we will first describe how we deal with insoluble graphs, in Section 5.4.3.1, then define a general model in Section 5.4.3.2.

5.4.3.1 Models for insoluble graphs

Certain graphs that are allowed by Theorem 10 violate solubility, and the constructions from Everitt et al. [2021a] and van Merwijk et al. [2022] will need to be altered in order to establish materiality in these graphs.

The assumption of solubility meant that upstream decisions could not contain latent, actionable information — in particular, this implied if an info path m_i contains a context C for a decision $X' \in \mathbf{X}(\mathcal{S}) \setminus \{X_i\}$, then V would have to be context of X_i , otherwise the past decision V would contain latent information that is of import to X_i [van Merwijk et al., 2022, Lemma 28]. For example, in Figure 5.7a the red info path contains the variable W_1 , which is a context for X' but not for X_0 , and solubility is violated because $W_1 \perp Y \mid \{Z_0, X_0, X_1\}$ but it satisfies all the three conditions of Theorem 10.

We can nonetheless apply the construction from [van Merwijk et al., 2022] to this graph, by treating X' as though it was a non-decision. This yields the

decision problem shown in Figure 5.7a, which is example of the construction from Figure 5.7c), except that there is a decision X' that observes Z_0 and W_1 . In this model, the outcome Y is equal to 1 if x_0 is equal to u_1 . The intended logic of this construction is that since $W_1 = Z_0 \oplus U_q$, the MEU can be achieved with the non-intervened policy $X_0 = Z_0 \oplus W_1$, which would require X_0 to depend on Z_0 . In this model, however, there exists an alternative policy where $X' = U_1$ and $X_0 = X'$, which achieves the MEU of 1, without having X_0 directly depend on Z_0 , and proving that Z_0 is immaterial for X_0 . Essentially, the single bit of X' sufficed to transmit the value of U_1 , meaning that Z_0 contained no more useful information. So long as the decision problem allows X' can do this there can be no need for X_0 to observe Z_0 . So in order to exhibit materiality, we need the domain of X' to be smaller than that of U_1 .

As such, we can devise a modified scheme, shown in Figure 5.7b. In this scheme, *two* random bits are generated at U_1 . The outcome is $Y = 1$ if X_1 supplies one bit from U_1 along with its index. A random bit is sampled at Z_0 , and W_1 presents the Z_0^{th} bit from U_1 , while X_1 has a domain of just one bit. Then, similar to our previous discussion of Figure 5.4b, the only bit from U_1 that X_0 can reliably know is the Z_0^{th} bit. Hence the only way achieve the MEU is for X' to inform X_0 about the value of W_1 , and for X_0 to equal $X_0 = \langle Z_0, X' \rangle$. Importantly, this can only be done if X_0 observes Z_0 ; it is material for X_0 .

In Figure 5.7b, if x_1 produces the z_0^{th} bit from u_1 , i.e. $x_1 = \langle z_0, u_1[z_0] \rangle$, we will call it *consistent* with $\langle z_0, u_1 \rangle$. If it produces *any* bit from u_1 , then we will call it *compatible* with $\langle z_0, u_1 \rangle$. For instance, either $\langle 0, 0 \rangle$ or $\langle 1, 1 \rangle$ is compatible with $z_0 = 0$ and $u_1 = 01$, but only the former is consistent with $z_0 = 0$ and $u_1 = 00$.

We can generalise these concepts to a case of multiple fork variables, rather than just Z_0 and U_1 . For example, Figure 5.7c, we have $J + 1$ fork variables $U_{0:J}$, which sample bitstrings of increasing length. Then, $Z_0 = W_u$, and each collider W_i has $W_i = U_j[U_{j-1}]$. The outcome Y will still check whether X_0 is compatible with U_J , but it will do so using a more general definition, as follows.

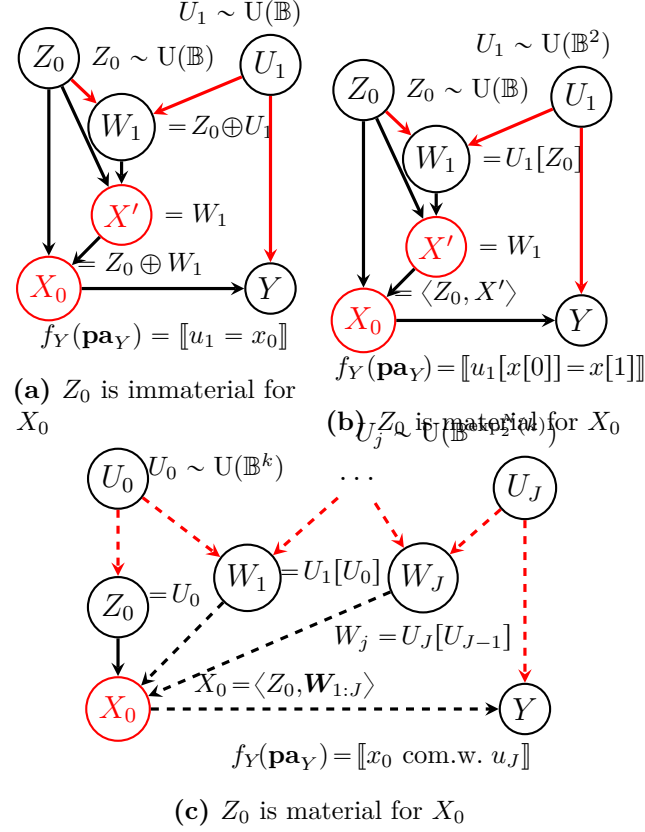


Figure 5.7: Two SCMs (a-b), and a description of a family of SCMs, where each dashed line represents a path. The repeated exponent $\exp_2^n(k)$ is defined as k if $n = 0$, and $2^{\exp_2^{n-1}(k)}$ otherwise.

Definition 31 (Consistency and compatibility). *Let $\mathbf{w} = \langle w_0, w_1, \dots, w_J \rangle$ where $w_0 \in \mathbb{B}^k$ and $w_n \in \mathbb{B}$ for $n \geq 1$. Then, \mathbf{w} is consistent with $\mathbf{u} = \langle u_0, \dots, u_J, u_i \in \mathbb{B}^{\exp_2^i(k)} \rangle$ (i.e. $\mathbf{w} \sim \mathbf{u}$) if $w_0 = u_0$ and $w_n = u_n[u_{n-1}]$ for $n \geq 1$. Moreover, \mathbf{w} is compatible with $u_J \in \mathbb{B}^{\exp_2^J(k)}$ (i.e. $\mathbf{w} \sim u_J$) if there exists any u_0, \dots, u_{J-1} such that \mathbf{w} is consistent with u_0, \dots, u_J .*

In Figure 5.7b, if, with positive probability, the assignment of X_0 is inconsistent with $\langle z_0, u_1 \rangle$, then the decision-maker is also penalised with strictly positive probability. For instance, if the assignments $z_0 = 0$ and $u_1 = 01$ lead to the assignment $x = \langle 1, 1 \rangle$, then this policy will achieve utility of $y = 0$ given the assignments $y_0 = 0$ and $u_1 = 00$, since they cause the values $z_0 = 0$ and $w_1 = 0$, which will cause the assignment $x = \langle 1, 1 \rangle$, which is not consistent with $z_0 = 0$ and $u_1 = \langle 0, 0 \rangle$. We find that the same is true in the more general mode of Figure 5.7c.

If with strictly positive probability, the assignment of X_0 is inconsistent with $\mathbf{u}_{0:J}$, then there will exist an alternative assignment $\mathbf{U}_{0:J} = \mathbf{u}'_{0:J}$, that produces the same assignments to the observations of X_0 , but where X_0 is not compatible with \mathbf{u}'_J .

Lemma 4. *Let $\mathbf{w} = \langle w_0, \dots, w_J \rangle$ and $\bar{\mathbf{w}} = \langle \bar{w}_0, \dots, \bar{w}_J \rangle$ be sequences with $w_0, \bar{w}_0 \in \mathbb{B}^k$, $w_j, \bar{w}_j \in \mathbb{B}$ for $j \geq 1$, and let $J' \leq J$ be the smallest integer such that $w_{J'} \neq \bar{w}_{J'}$. Let $u_0, \dots, u_{J'}$ be a sequence where $u_j[u_{j-1}] = w_j$ for $1 \leq j < J'$. Then, there exists some $u_{J'+1}, \dots, u_J$ such that \mathbf{w} is consistent with u_0, \dots, u_J , but $\bar{\mathbf{w}}$ is incompatible with u_J .*

The proof is deferred to Appendix C.2.5.

This result implies that an optimal policy in Figure 5.7c, x_0 must be consistent with $\mathbf{u}_{0:J}$ with probability 1. After all, the non-intervened policy clearly achieves the MEU of 1, being that it is consistent with $\mathbf{u}_{0:J}$, and consistency implies compatibility. On the other hand, if x_0 is inconsistent with $\mathbf{u}_{0:J}$ with strictly positive probability, then there will exist an alternative assignment $\mathbf{u}'_{0:J}$ that produces the same assignment x_0 , and since the variables $\mathbf{U}_{0:J}$ have full support, this will lead to $y = 0$ with strictly positive probability, and decrease the expected utility. If a policy cannot copy Z_0 without observing it, then this will make X_0 inconsistent with \mathbf{u} with strictly positive probability, so this policy will not be optimal. One may notice that by setting U_0 to contain k bits rather than just one, this will make it very difficult for Z_0 to copy the value of Z_0 without observing it, if a sufficiently large k is chosen. We will develop a fully formal argument for materiality in Section 5.4.4.

5.4.3.2 A decision problem for any graph containing the materiality paths

We will now generalise the constructions from Figure 5.3a (for a truncated info path is a directed path) and Figure 5.7c (for a truncated info path that is not a directed path) to an arbitrary graph containing the materiality paths described in Lemma 3.

To begin with, let us note that the materiality paths may overlap. So our general approach will be to define a random variable V^p for each variable in a path p . To

derive the overall materiality SCM, we will simply define V by a cartesian product over each V_p . For the outcome variable Y , we will instead take a sum over each Y^p . For any set of paths \mathbf{p} , we define $V^{\mathbf{p}} = \times_{p \in \mathbf{p}} V^p$.

Let us now discuss the control path. The initial node A will sample a bitstring that is passed along the control path, and through each intersection node T_i in particular. To describe this, we will rely on shorthand.

Definition 32 (Parents along paths). *When a vertex V has a unique parent \bar{V} along p , $\mathbf{Pa}(V^p) = \bar{V}^p$, and for a set of paths \mathbf{p}' , let $\mathbf{Pa}(V^{\mathbf{p}'}) = \times_{p \in \mathbf{p}'} \mathbf{Pa}(V^p)$. For a collider V in a truncated info path $m_i : T_i \text{ -- } Y$, let the parent nearer T_i along m_i be $\mathbf{Pa}_L(V)$, and the parent nearer Y be $\mathbf{Pa}_R(V)$.*

For example, a non-outcome child V of A along the control path will be assigned $V^d = \mathbf{Pa}(V^d)$.

Each info path must pass on information from upstream paths that traverse the intersection node. We therefore use the notation \mathbf{p}_i to refer to the set of control and auxiliary paths that enter the intersection node T_i . We also devise an extended notion of parents \mathbf{Pa}^* to include this information. Relatedly, we will define a notion of parents for the auxiliary path, which includes information from the collider $W_{i,j}$ of the info path, and a notion of parents for the paths \mathbf{p}_i , that includes the exogenous parent \mathcal{E}_A of A .

Definition 33 (Extended parent relations). *For a truncated info path m_i , let:*

$$\mathbf{Pa}^*(V^{m_i}) = \begin{cases} T_i^{\mathbf{p}_i} & \text{if } \mathbf{Pa}(V^{m_i}) = T_i^{m_i} \\ \mathbf{Pa}(V^{m_i}) & \text{otherwise} \end{cases} \quad \text{and}$$

$$\mathbf{Pa}_l^*(V) = \begin{cases} T_i^{\mathbf{p}_i} & \text{if } \mathbf{Pa}_L(V^{m_i}) = T_i^{m_i} \\ \mathbf{Pa}_L(V^{m_i}) & \text{otherwise} \end{cases}.$$

$$\text{For an auxiliary path } r_{i,j}, \text{ let } \mathbf{Pa}^*(V^{r_{i,j}}) = \begin{cases} W_{i,j}^{m_i} & \text{if } \mathbf{Pa}(V^{r_{i,j}}) = W_{i,j}^{m_i} \\ \mathbf{Pa}(V^{r_{i,j}}) & \text{otherwise} \end{cases}.$$

$$\text{Finally, let: } \mathbf{Pa}^*(V^{\mathbf{p}_i}) = \begin{cases} \mathcal{E}_A \times \mathbf{Pa}(V^{\mathbf{p}_i}) & \text{if } V \text{ is } A \\ \mathbf{Pa}(V^{\mathbf{p}_i}) & \text{otherwise} \end{cases}.$$

In other respects, the materiality SCM will behave in a similar manner to previous examples. For instance, when m_i is directed, the outcome Y^{m_i} will evaluate whether

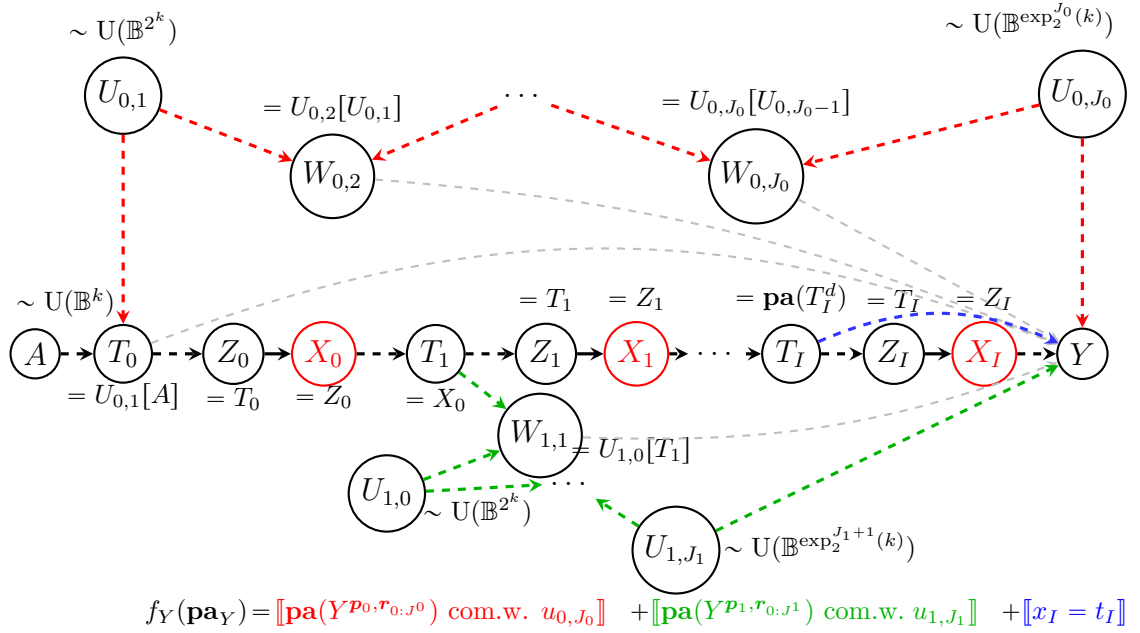


Figure 5.8: The materiality SCM: a general SCM where Z_0 is material for X_0 .

the values $\mathbf{Pa}(Y^{p_i})$ (which mostly come from X_i) are equal to $\mathbf{Pa}(Y^{m_i})$, which come from the info path. When m_i is not directed, the outcome Y^{m_i} will evaluate whether the values from $\mathbf{Pa}(Y^{p_i, r_{i,0:J}})$ are compatible with those from $U_{i,J}$. So let us now define the materiality SCM as follows.

Definition 34 (Materiality SCM). *Given a graph containing the materiality paths, we may define the following random variables.*

In the control path, $d : A \dashrightarrow Y$, let:

- *the source be $A^d = \mathcal{E}^{A^d}$ where $\mathcal{E}^{A^d} \sim U(\mathbb{B}^k)$ where k is the smallest positive integer such that $2^k > (k+c)bc$, where b is the maximum number of variables that are contexts of one decision, $b := \max_{X \in \mathbf{X}(S)} |C_X|$, and c is the maximum number of materiality paths passing through any vertex in the graph;*
- *every non-endpoint V have $V^d = \mathbf{Pa}(V^d)$.*

In each truncated info path that is directed, $m_i : T_i \dashrightarrow Y$, let:

- *the intersection node T^{m_i} have trivial domain;*

- each chain node be $V^{m_i} = \mathbf{Pa}^*(V^{m_i})$
- the outcome have the function $f_{Y^{m_i}}(\mathbf{pa}_Y) = \llbracket \mathbf{pa}(Y^{p_i}) = \mathbf{pa}^*(Y^{m_i}) \rrbracket$.

In each truncated info path that is not directed, $T_i \dashleftarrow \leftarrow W_{i,1} \rightarrow \cdots \leftarrow W_{i,J} \dashrightarrow Y$, let:

- each fork be $W_{i,j}^{m_i} = \mathcal{E}^{W_{i,j}^{m_i}}, \mathcal{E}^{W_{i,j}^{m_i}} \sim U(\mathbb{B}^{\exp_2^j(k+|\mathbf{p}_i|-1)})$ where $|\mathbf{p}_i|$ is the number of paths in \mathbf{p}_i ;
- each chain node be $V^d = \mathbf{Pa}^*(V^d)$;
- each collider be $V^{m_i} = \mathbf{Pa}_R(V^{m_i})[\mathbf{Pa}_L^*(V^{m_i})]$;
- each intersection node be $T_i^{m_i} = \mathbf{Pa}(V^{m_i})[\mathbf{Pa}^*(T_i^{p_i})]$ if the info path begins as $T_i \rightarrow \cdot$, otherwise it has empty domain;
- the outcome have the function $f_{Y^{m_i}}(\mathbf{pa}_Y) = \llbracket \mathbf{pa}(Y^{p_i, r_{i,1:J_i}}) \text{ is compatible with } \mathbf{pa}^*(Y) \rrbracket$.

In each auxiliary path $r_{i,j} : W_{i,j} \rightarrow V_2 \dashrightarrow Y$, let:

- each chain node have $V^{r_{i,j}} = \mathbf{Pa}^*(v^{r_{i,j}})$.
- each source $W_{i,j}$ have trivial domain

Then, let the materiality SCM have outcome variable $Y = \sum_{i_{\min} \leq i \leq i_{\max}} Y^{m_i}$, and non-outcome variables $V = \times_{p \in \{d, m_i, r_{i,1:J_i} \mid i_{\min} \leq i \leq i_{\max}\}} V^p$.

Note that this defines an SCM because each variable is a deterministic function of only its endogenous parents and exogenous variables.

We have define the materiality SCM so that decisions behave just as non-decisions, which always do what is required to ensure that $Y^{m_i} = 1$.

Lemma 5. *In the non-intervened model, the materiality SCM has $Y = i_{\max} - i_{\min} + 1$, surely.*

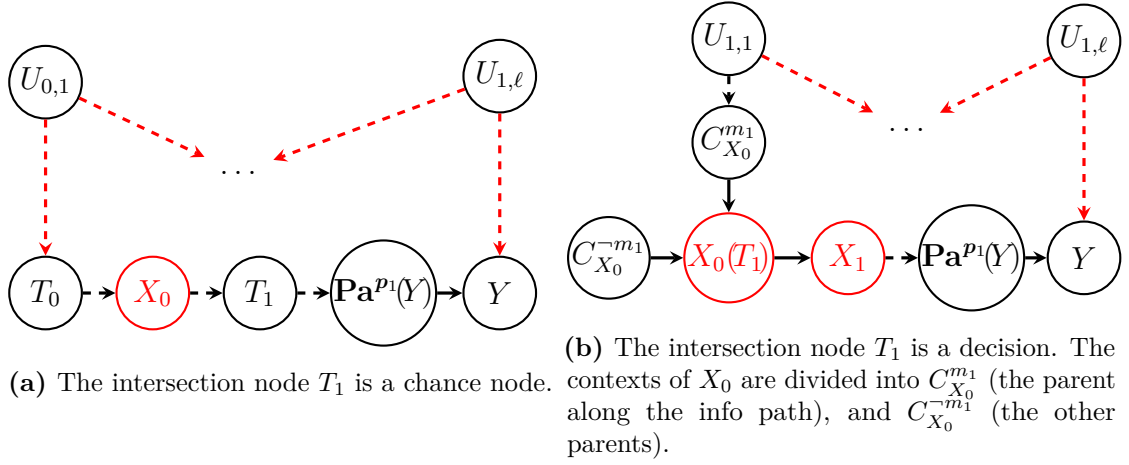


Figure 5.9: The cases where the intersection node T_1 is a chance node, or a decision

The proof follows from the model definition, and is supplied in Appendix C.2.4.

We also know that each utility term Y^{m_i} is upper bounded at one, so in order to obtain the MEU, each Y^i must equal 1, almost surely.

Lemma 6. *If a policy π for the materiality SCM, has $P^\pi(Y^{m_i} < 1) > 0$ for any i , the MEU is not achieved.*

Proof. We know that $\mathbb{E}^\pi[Y] = \sum_{i_{\min} \leq i \leq i_{\max}} Y^{m_i}$ (Definition 34), so for all i , $Y^{m_i} \leq 1$ always. So, if $P^\pi(Y^{m_i} < 1) > 0$ for any i , then $\mathbb{E}^\pi[Y] < i_{\max} - i_{\min} + 1$, which underperforms the policy that is followed in the non-intervened model (Lemma 5). □

5.4.4 Proving materiality in the materiality SCM

We will now prove that in the materiality SCM, if Z_0 is removed from the contexts of X_0 , then the performance for at least one of the utility variables Y^{m_i} is compromised, and so the MEU is not achieved. The proof divides into two cases, based on whether the child of X_0 along the control path is a non-decision (Section 5.4.4.1) or a decision (Section 5.4.4.2).

5.4.4.1 Case 1: child of X_0 along d is a non-decision.

If the child of X_0 along the control path is a non-decision and Z_0 is not a context of X_0 , we will prove that $\mathbb{E}[Y^{m_0}] < 1$. In this case, either X_0 is the last decision in the control path, or otherwise there must exist an intersection node T_1 , as shown in Figure 5.9a. If the former is true, then it is immediate that the value x_0 is transmitted to Y along the control path, based on the model definition. As such, Y_0 can directly evaluate the decision X_0 . For the latter case, we want an assurance that downstream decisions will pass along the value of X , as was the case in Figure 5.4b. Such an assurance is provided by the following lemma, which shows that whenever an intersection node T_i is a chance node — as is T_1 — the value t_i is transmitted to Y by every optimal policy.

Lemma 7 (Chance intersection node requirement). *If in the materiality SCM, where T_i is a chance node, a policy π has $P^\pi(\mathbf{Pa}(T_i^{p_i}) = \mathbf{Pa}(Y^{p_i})) < 1$, then $P^\pi(Y^{m_i} < 1) > 0$.*

First, we prove the case where m_i is a directed path. In this case, m_i copies the value t^{p_i} to Y , which Y^{m_i} checks against the value $\mathbf{pa}(y^{p_i})$ received via the control path. Maximising Y^{m_i} then requires them to be equal.

Proof of Lemma 7 when m_i is a directed path. We have $f_{Y^{m_i}}(\mathbf{pa}_{Y^{m_i}}) = \llbracket \mathbf{pa}(Y^{m_i}) = \mathbf{pa}(Y^{p_i}) \rrbracket$ (Definition 34). Also, $\mathbf{Pa}(Y^{m_i}) = T_i^{p_i} = \mathbf{Pa}(T_i^{p_i})$ surely, where the first equality follows from Definition 34, while the second follows from Definition 34 and T_i being a chance node. So, if $P^\pi(\mathbf{Pa}(Y^{p_i}) = \mathbf{Pa}(T_i^{p_i})) < 1$ then $P^\pi(Y^{m_i} = 1) < 1$. \square

We now prove the case where m_i is a directed path. In this case, if the assignment $\mathbf{pa}(Y^{p_i})$ transmitted along the control path differs from the value $\mathbf{pa}(T_i^{p_i})$ that came in to the intersection node T_i , then just as we established for Figure 5.7c, there will exist an assignment $\mathbf{u}_{i,1:J_i}$ to the fork nodes in m_i that gives an unchanged assignment to colliders $\mathbf{v}_{i,1:J_i}$, but where $\mathbf{pa}(Y^{p_i})$ is incompatible with u_{J_i} .

Proof of Lemma 7 when m_i is not a directed path. Let us index the forks and colliders of m_i as $T_i \dashleftarrow V_{i,1} \dashleftarrow U_{i,1} \dashrightarrow W_{i,1} \dashleftarrow \dots W_{i,J_i} \dashleftarrow U_{i,J_i} \dashrightarrow Y$. Choose any assignments $\mathbf{pa}(T_i^{p_i}) \neq \mathbf{pa}(Y^{p_i})$ that occur with strictly positive probability. Then, there must also exist assignments $\mathbf{Pa}(Y^{p_i, r_{i,1}:J_i}) = \mathbf{pa}(Y^{p_i, r_{i,1}:J_i})$, $\mathbf{U}_{i,1:J_i} = \mathbf{u}_{1:J_i}$, and $\mathbf{W}_{i,1:J_i} = \mathbf{w}_{1:J_i}$ such that

$$P^\pi(\mathbf{pa}(T_i^{p_i}), \mathbf{pa}(Y^{p_i, r_{i,1}:J_i}), t_i^{p_i}, \mathbf{u}_{1:J_i}, \mathbf{w}_{1:J_i}) > 0.$$

By Lemma 4, there also exists an assignment $\mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i}$ such that $\mathbf{pa}(T_i^{p_i}), \mathbf{w}_{1:J_i}$ is consistent with $\mathbf{u}'_{1:J_i}$, and $\mathbf{pa}(Y_i^p), \mathbf{pa}(Y^{r_{i,1}:J_i})$ is incompatible with \mathbf{u}'_{J_i} . Now, consider the intervention $\text{do}(\mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i})$. Since T_i is a chance node, every collider in m_i is a non-decision, and is assigned the (unique) value consistent with $\mathbf{pa}(T_i^{p_i}), \mathbf{u}'_{1:J_i}$. Furthermore, $\mathbf{pa}(T_i^{p_i}), \mathbf{w}_{1:J_i}$ is consistent with $\mathbf{pa}(T_i^{p_i}), \mathbf{u}'_{1:J_i}$, so the intervention does not affect the assignments to these colliders. Moreover, from Definition 34, no variable outside of m_i is affected by assignments within m_i , except through the colliders. Therefore:

$$\begin{aligned} & P^\pi(\mathbf{pa}(Y^{p_i}), \mathbf{pa}(Y^{r_{i,1}:J_i}), \mathbf{Pa}(Y^{m_i}) = u'_{J_i} \mid \text{do}(\mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i})) > 0 \\ \therefore & P^\pi(Y^{m_i} = 0 \mid \text{do}(\mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i})) > 0 \\ & (\mathbf{pa}(Y_i^p), \mathbf{pa}(Y^{r_{i,1}:J_i})) \text{ not compatible with } \mathbf{u}'_{J_i} \\ \therefore & P^\pi(Y^{m_i} = 0 \mid \mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i}) > 0 \\ & (\mathbf{U}_{i,1:J_i} \text{ are unconfounded, so } P^\pi(\mathbf{V} \mid \text{do}(\mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i})) = P^\pi(\mathbf{V} \mid \mathbf{U}_{i,1:J_i} = \mathbf{u}'_{1:J_i})) \\ \therefore & P^\pi(Y^{m_i} = 0) > 0 \qquad (P^\pi(\mathbf{u}_{i,1:J_i}) > 0). \end{aligned}$$

□

If m_i is not a directed path, then this requirement extends to the values $\mathbf{pa}(Y^{r_{i,1}:J_i})$ passed down the auxiliary paths, not just the value $\mathbf{pa}(Y^{p_i})$ from the control path. Specifically, $\mathbf{pa}(Y^{p_i}), \mathbf{pa}(Y^{r_{i,1}:J_i})$ must be consistent with $\mathbf{pa}(Y^{p_i}), \mathbf{u}_{i,1:J_i}$, where $\mathbf{u}_{i,1:J_i}$ denotes the values of forks on the info path.

Lemma 8 (Collider path requirement). *If the materiality SCM has an info path m_i that is not directed, and under the policy π there are assignments $\mathbf{Pa}(Y^{p_i, r_{i,1:j_i}}) = \mathbf{pa}(Y^{p_i, r_{i,1:j_i}})$ to parents of the outcome, and $U_{i,1:j_i}^{m_i} = \mathbf{u}_{i,1:j_i}^{m_i}$ to the forks of m_i , with $P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:j_i}}), \mathbf{u}_{i,1:j_i}^{m_i}) > 0$ and where $\mathbf{pa}(Y^{p_i, r_{i,1:j_i}})$ is inconsistent with $\mathbf{pa}(Y^{p_i}), \mathbf{u}_{i,1:j_i}^{m_i}$, then $P^\pi(Y^{m_i} < 1) > 0$.*

The idea of the proof, similar to Lemma 7, is that whenever the bits transmitted along the auxiliary paths deviate from the values $\mathbf{w}_{i,1:j_i}$ of colliders in m_i , there exists an assignment $\mathbf{u}'_{i,1:j_i}$ to forks in m_i that will render the colliders, and hence the decision x_i unchanged, while making x_i incompatible with u_{j_i} , and thereby producing $Y^{m_i} < 0$. A detailed proof is in Appendix C.2.5.

In order to prove that the context Z_0 is needed, we will also need to establish that it is not deterministic, even if it is a decision. In the case where Z_0 is a decision, the idea is that random information is generated at A , which each of the decisions are required to pass along the control path. We are able to prove this as a corollary of Lemma 7.

Lemma 9 (Initial truncated info path requirements). *If π in the materiality SCM does not satisfy: $P^\pi(\mathbf{Pa}(Y^d) = A^d) < 1$. then the MEU is not achieved.*

Proof. From Lemma 3, the control path d begins with a chance node. So, the first decision $X_{i_{\min}}$ in d must have a chance node $Z_{i_{\min}}$ as its parent along d . Furthermore, the intersection node $T_{i_{\min}}$ must be an ancestor of $Z_{i_{\min}}$ along d , so it is also a chance node. So it follows from Lemma 7, that any policy π must satisfy $P^\pi(T_{i_{\min}}^{p_{i_{\min}}} = \mathbf{Pa}(Y^{p_{i_{\min}}})) = 1$ if it attains the MEU. As $T_{i_{\min}}$ is in the control path, we have $d \in \mathbf{p}_{i_{\min}}$ (Lemma 3) so $T_{i_{\min}}^d \stackrel{\text{a.s.}}{=} \mathbf{Pa}(Y^d)$ is also required. Moreover, all of vertices in the segment $A \dashrightarrow T_{i_{\min}}$ of d are chance nodes, because $X_{i_{\min}}$ was defined as the first decision in d , and $T_{i_{\min}}$ precedes it. And, each chance variable V^d on the control path equals its parent $\mathbf{Pa}(V^d)$ (Definition 34), so $A^d = T_{i_{\min}}^d$, and thus $A^d \stackrel{\text{a.s.}}{=} \mathbf{Pa}(Y^d)$ is required to attain the MEU. \square

We can now combine our previous results to prove that it is impossible to achieve the MEU, if Z_0 is not a context of X_0 , in the case where T_1 does not exist, or is a non-decision.

Lemma 10 (Required properties unachievable if child is a non-decision). *Let \mathcal{M} be a materiality SCM where the child of X_0 along d is a non-decision. Then, the MEU for the scope \mathcal{S} cannot be achieved by a deterministic policy in the scope $\mathcal{S}_{Z_0 \rightarrow X_0}$ (equal to \mathcal{S} , except that Z_0 is removed from \mathbf{C}_{X_0}).*

The logic is that if child of X_0 in the control path is a non-decision, then the value of X_0 is copied all the way to $\mathbf{Pa}(Y^d)$ (Lemma 9). Furthermore, $Z_0^d \stackrel{\text{a.s.}}{=} \mathbf{Pa}(Y^d)$ is necessary to achieve the MEU (Lemma 7). But the materiality SCM has been constructed so that the non- Z_0 parents of X_0 do not contain enough bits to transmit all of the information about Z_0^d , so the MEU cannot be achieved. The proof is detailed in Appendix C.2.6.

5.4.4.2 Case 2: child of X_0 along d is a decision.

If the child of X_0 along d is a decision, as shown in Figure 5.9b, we will prove that the decision X_0 must depend on Z_0 in order to achieve $\mathbb{E}[Y_1] = 1$. This will be because without Z_0 , X_0 will be limited in its ability to distinguish all of the possible values of the first fork node $U_{i,1}$ of m_1 . To establish this, we will need to conceive of a possible intervention to the fork nodes in m_i , that X_i would have to respond to, and so we begin by proving that relatively few variables will be causally affected by certain interventions.

Lemma 11 (Fork information can pass in few ways). *If, in the materiality SCM:*

- *the intersection node T_i is the vertex X_{i-1} ,*
- *π_{T_i} is a deterministic decision rule where $\pi_{T_i}(\mathbf{c}^{\neg m_i}(T_i, u_{i,1})) = \pi_{T_i}(\mathbf{c}^{\neg m_i}(T_i, u'_{i,1}))$ for assignments $u_{i,1}, u'_{i,1}$ to the first fork variable, and $\mathbf{c}^{\neg m_i}(T_i)$ to the contexts of T_i not on m_i , and*

- $\mathbf{W}_{i,1:J_i} = \mathbf{w}_{i,1:J_i}$, and $\mathbf{U}_{i,2:J_i} = \mathbf{u}_{i,2:J_i}$ are assignments to forks and colliders in m_i where each $u_{i,j}$ consists of just $w_{i,j}$ repeated $\exp_2^j(k + |\mathbf{p}_i| - 1)$ times, then:

$$\begin{aligned} & P^\pi(\mathbf{pa}(Y^{\mathbf{p}_i, r_{i,1}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{i,1:J_i}, \mathbf{u}_{i,2:J_i} \mid \text{do}(u_{i,1})) \\ &= P^\pi(\mathbf{pa}(Y^{\mathbf{p}_i, r_{i,1}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{i,1:J_i}, \mathbf{u}_{i,2:J_i} \mid \text{do}(u'_{i,1})). \end{aligned}$$

The proof follows from the definition of the materiality SCM, and it is detailed in Appendix C.2.7.

We can now prove that if a deterministic policy does not appropriately distinguish assignments to $U_{i,1}$, then the i^{th} component of the utility will be sub-optimal $\mathbb{E}[Y^{m_i}] < 1$.

Lemma 12 (Decision must distinguish fork values). *If in the materiality SCM:*

- the intersection node T_i is the vertex X_{i-1} , and
 - π is a deterministic policy that for assignments $u_{i,1}, u'_{i,1}$ to $U_{i,1}$ where $u_{i,1} \neq u'_{i,1}$, has $\pi_{T_i}(\mathbf{c}^{-m_i}(T_i), u_{i,1}) = \pi_{T_i}(\mathbf{c}^{-m_i}(T_i), u'_{i,1})$ for every $\mathbf{C}_{T_i}^{-m_i}(T_i) = \mathbf{c}^{-m_i}(T_i)$,
- (†)

then $P^\pi(Y^{m_i} < 1) > 0$

The idea of the proof is that if $u_{i,1}$ and $u'_{i,1}$ differ, there will be some assignment $\mathbf{pa}(Y^{\mathbf{p}_i})$ such that $u_{i,1}[\mathbf{pa}(Y^{\mathbf{p}_i})]$ and $u'_{i,1}[\mathbf{pa}(Y^{\mathbf{p}_i})]$ differ. When $\mathbf{Pa}(Y^{\mathbf{p}_i}) = \mathbf{pa}(Y^{\mathbf{p}_i})$ and $u_{i,1}$, then $\mathbf{Pa}(Y^{r_{i,1}})$ to assume one value. But if we intervene $u'_{i,1}, u_{i,2:J_i}$, then the value of $\mathbf{Pa}(Y^{r_{i,1}})$ will be incorrect, making $\mathbf{Pa}(Y^{\mathbf{p}_i, r_{i,1:J_i}})$ inconsistent with $\mathbf{Pa}(Y^{\mathbf{p}_i}, U_{i,1:J_i})$ so the maximum expected utility is impossible to achieve. The details are deferred to Appendix C.2.8.

This will allow us to prove that when the child of X_0 along d is a decision, the MEU cannot be achieved without Z_0 as a context of X_0 .

Lemma 13 (Required properties unachievable if child is a decision). *Let \mathcal{M} be the materiality SCM for some scoped graph \mathcal{G}_S , where $i_{\max} > 0$ and T_1 is a decision. Then, there exists no deterministic policy in the scope $\mathcal{S}_{Z_0 \rightarrow X_0}$ that achieves the MEU.*

To prove that no deterministic policy in $\mathcal{S}_{Z_0 \rightarrow X_0}$ can achieve the MEU (achievable with the scope \mathcal{S}), we will show that if a deterministic policy π satisfies $P^\pi(\mathbf{Pa}(Y^d) = A^d) = 1$, as required by Lemma 9, then the domain of $X_0 \times C_{X_0}^{-m_1}$ is smaller than the domain of $C_{X_0}^{m_1}$, so Equation (†) will be satisfied, and thus the MEU cannot be achieved. A detailed proof is presented in Appendix C.3.

We now combine the lemmas for the two cases to prove the main result.

Proof of Theorem 10. Any scoped graph $\mathcal{G}(\mathcal{S})$ that satisfies assumptions (A-C) contains materiality paths for the context Z_0 of X_0 (Lemma 3), and has a materiality SCM (Definition 34) compatible with $\mathcal{G}(\mathcal{S})$. In this decision problem, whether the child of X_0 along d is or is not a decision, the MEU cannot be achieved by a deterministic policy unless X_0 is allowed to depend on Z_0 (Lemmas 10 and 13). And stochastic policies can never surpass the best deterministic policy ([Lee and Bareinboim, 2020, Proposition 1]), so no such policy can achieve the MEU, and so Z_0 is material for X_0 . □

5.5 Toward a more general proof of materiality

So far, via Theorem 10 we have established the necessity of condition (I) of LB-factorizability for immateriality. We now outline some steps toward evaluating the necessity of conditions (II-III) of LB-factorizability, and the further condition in [Lee and Bareinboim, 2020, Thm. 2].

To begin with, condition (III) requires that we choose an ordering $<$, such that the parents of each decision X are somewhere before X , while the descendants are somewhere afterwards. Clearly this condition can be satisfied for any acyclic graph, so it instead

Conditions (II-III) are individually not very restrictive, but are jointly substantial. So a natural next step is to try to prove that conditions (II-III) are necessary, by defining some info paths and control paths for graphs that violate conditions (II-III), defining a materiality SCM, and proving materiality in that SCM. So far,

however, we have only been able to carry out the first step — defining the paths — and difficulties have arisen in using those paths to define an SCM that exhibits materiality. In this section, we will outline what info paths and control paths can be proven to exist, and then outline the difficulties in using them to prove materiality.

5.5.1 A lemma for proving the existence of paths

When the variables $\mathbf{Z}, \mathbf{X}', \mathbf{C}', \mathbf{U}$ are not factorizable, we can prove the existence of info and control paths.

Lemma 14 (System Exists General). *Let \mathcal{G}_S be a scoped graph that satisfies assumptions (A, B) from Theorem 10. If $\mathbf{Z} = \{Z_0\}$, $\mathbf{X}' \supseteq \mathbf{Ch}(Z_0)$, $\mathbf{C}' = C_{\mathbf{X}'} \setminus (\mathbf{X}' \cup \mathbf{Z})$, $\mathbf{U} = \emptyset$ are not LB-factorizable, then there exists a pair of paths to some $C' \in \mathbf{C}' \cup Y$:*

- an info path $m : Z_0 \dashrightarrow C'$, active given $[\mathbf{X}' \cup \mathbf{C}']$, and
- a control path $d : X \dashrightarrow C'$ where $X \in \mathbf{X}'$.

A proof is supplied in Appendix C.4.1. The intuition of this proof is that each of the conditions (I-III) implies a precedence relation between a pair of variables in $\mathbf{V}' \cup Y$. Each of these precedence relations can be used to build an “ordering graph” over $\mathbf{V}' \cup Y$. If the ordering graph is acyclic, then we can let $<$ be any ordering that is topological on the graph, and then $\mathbf{Z}, \mathbf{X}', \mathbf{C}', \mathbf{U}$ are LB-factorizable. Otherwise, we can use a cycle in the graph to prove the existence of an info path and a control path. By iterating through these cycles, we can obtain a series of info paths and control paths that terminate at Y .

The resulting paths are in some cases, quite useful for proving materiality. For instance, we can recover the pair of info and control paths used in Figure 5.4b. To prove that Z is material for X , we can start by choosing $\mathbf{X}' = \{X, X'\}$, $\mathbf{C}' = \{Z'\}$, $\mathbf{C}' = \{Z', W\}$, and $\mathbf{U}' = \emptyset$. Then, Lemma 14 implies the existence of an active path from Z to some $\mathbf{Desc}_X \cap \mathbf{C}'$, so we see that the first info path is the edge $Z \rightarrow Z'$. With Z' being a descendant of X , we also have the first control path,

$X \rightarrow Z'$. We must then obtain some paths that exhibit why Z' is itself useful for the decision X to know about, and to influence. To do this, we can reapply Lemma 14 using the sets $\mathbf{X}' = \{X'\}$, $\mathbf{Z} = \{Z'\}$, $\mathbf{C}' = \{W\}$, and $\mathbf{U}' = \emptyset$. We then obtain the new info path $Z' \rightarrow W \leftarrow U \rightarrow Y$, and the new control path $Z' \rightarrow X' \rightarrow Y$. The SCM in Figure 5.4b uses these paths to prove Z material for X .

5.5.2 A further challenge: non-collider contexts

In some graphs, it is not clear how to use the info and control paths Lemma 14 to prove materiality, because non-collider nodes on the info path may be contexts. (In previous work, this possibility was excluded by the solubility assumption [van Merwijk et al., 2022, Lemma 28].) We will now highlight one case, in Figure 5.10, where it is relatively clear how this challenge can be overcome, and one case, Figure 5.11, where it is unclear how to make progress.

In the graph of Figure 5.10, we would like to prove that Z_0 is material for X_0 . Using Lemma 14, we can obtain the red and blue info paths as shown, and the corresponding control paths in darker versions of the same colors. In the approach of Definition 34, shown in Figure 5.10a, X_0 should need to observe Z_0 in order to know which slice from V is presented at its parent X_1 . Then, X_1 would play two roles, one for the red info path, and one for the dark blue control path. As a collider on the red info path, its role is to present the Z_0^{th} bit from V . As the initial endpoint of the blue control path, so its role is to copy the assignment of Z_0 . The problem, however, is that X_0 then does not need to observe Z_0 in order to reproduce its value, because this value is already observed at X_1 , so Z_0 is not material.

To remedy this problem, we can construct an alternative SCM, where the value of Z_0 is “concealed”, i.e. it is removed from the other contexts, $C_{Z_0} \setminus Z_0$. At X_1 , we directly remove Z_0 , leaving this decision with a domain of only one bit. At C , we impose some random noise, so that it is not always a perfect copy of Z_0 . The result is shown in Figure 5.10b. When this model is not intervened, an expected utility of $\mathbb{E}[Y] = 10.99$ is achieved, because the red term in Y always equals 10, while the

blue term has an expectation of 0.99. (This is the MEU, because there is no way to improve the blue term to have expectation 1 without decreasing the expectation of the red term by at least 0.05.) If instead, Z_0 is removed as a context for X_0 , then the expected utility can only be as high as $\mathbb{E}[Y] = 10.95$. To understand this, restrict our attention to deterministic policies, and note that in order for the red term to be better than a coin flip (with an expected value of 5), we would either need to have $X_0 = \langle C, X_1 \rangle$ — and the red term will have an expectation of 9.95, or we must have $X_1 = V[0]$ and $X_0 = \langle 0, X_1 \rangle$ — and then the blue term will have an expectation of 0.5. In either case, performance is worse than 10.99, so Z_0 is material for X_0 .

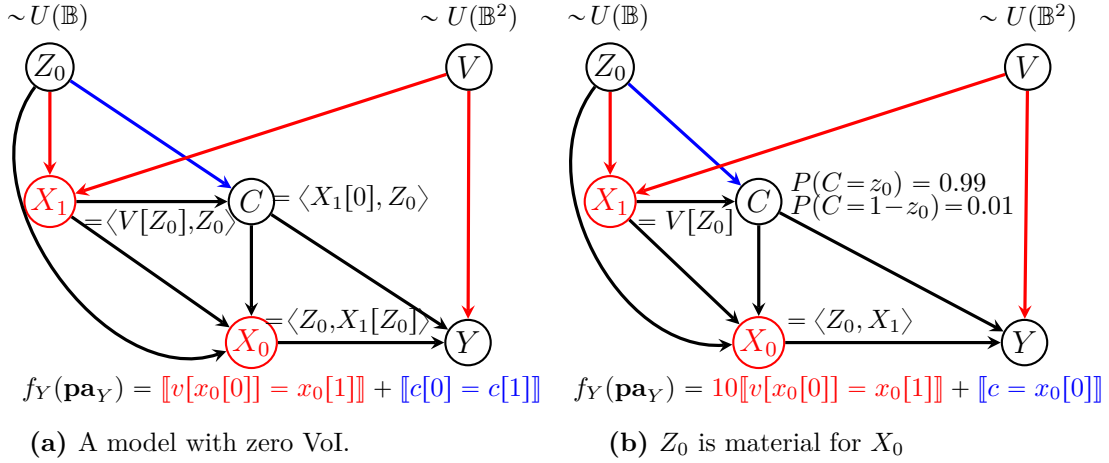


Figure 5.10: Two alternative models that use the same two info paths, red and blue.

The problem is that concealing the value of Z_0 does not work for all graphs. To see this, let us add two decisions, X_2 and X_3 , to the graph from Figure 5.10, to thereby obtain the graph in Figure 5.11. Let us retain the materiality SCM from Figure 5.10b, except that X_2 and X_3 copy the value from C along to Y . One might expect that Z_0 should still be material, but it is not. Now, there is a policy that achieves the new MEU of 11 by superimposing the value of Z_0 on the assignments of decisions X_2 and X_3 . In this policy π , $x_1 = v[z_0]$, $x_2 = z_0 \oplus z_0$, $x_3 = x_2 \oplus z_0$, and $x_0 = x_2 \oplus x_3 = z_0$ where \oplus represents the XOR function. Under π , the red term equals 10 always, while the blue term always equals 1, i.e. the MEU is achieved, and π is a valid policy even if Z_0 is not a context of

X_0 , meaning that Z_0 is not material for X_0 .

In summary, whenever $\mathbf{Z} \ni Z_0, \mathbf{X}' \ni X_0, \mathbf{C}', \mathbf{U}$ are not LB-factorizable, then we can find some info and control paths for Z_0 and X_0 , but then X_0 can recover the value of Z_0 , making it possible to achieve the MEU even when Z_0 is removed as a context of X_0 . In some graphs, we can devise an alternative SCM that conceals the value of Z_0 . But in others, a policy can superimpose the information from Z_0 on other decisions, such as X_2 and X_3 in Figure 5.11, so that X_0 can recover the value of Z_0 , making Z_0 immaterial for X_0 once again.

It seems that new insights are needed to solve this superimposition problem, and that therefore that we will need new insights to establish a complete criterion for materiality in insoluble decision problems.

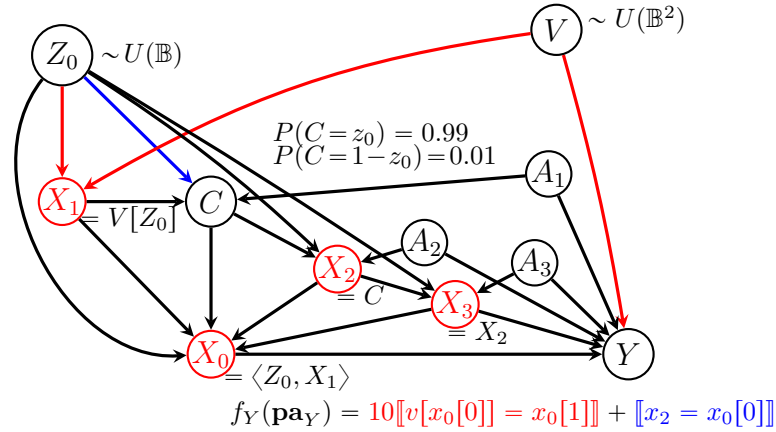


Figure 5.11: A model with zero Vol

5.6 Conclusion

We have found that in a graph whose contexts cannot satisfy condition (I) of LB-factorizability, any context can be material. We encountered some new problems for materiality proofs, and devised appropriate solutions:

- if the variable Z_i whose materiality we are trying to establish is a decision, whose value can be determined by other available contexts, — then we must

choose a different info path so that non-observed variables would be needed to determine the value of Z_i

- if the info path begins with a context of multiple decisions, — then we must construct the SCM differently along the info path
- if the control path contains consecutive decisions, then we require more bits to be copied along the control path, so that not all of these bits can be copied along alternative paths.

As a next step towards establishing a complete criterion for materiality, we then considered the more general setting where no context can jointly satisfy conditions (I-III) of LB-factorizability. In this setting, it is possible to identify info paths and control paths for a target context Z_0 and decision X_0 , and to apply our SCM construction to these paths. However, there may exist policies that transmit the assignment of Z_0 through alternative paths, and that achieve the MEU even when Z_0 is removed as a context of X_0 . Although there exist ways of concealing the information about Z_0 from a descendant decision $X_{i'}, i < i'$, there can also be other ways that information about Z_0 may be transmitted, such as transmitting this information in other decisions, undermining materiality once again. Thus, the challenge of proving a complete criterion of materiality for insoluble graphs currently remains open.

5.7 Acknowledgements

Thanks to Minwoo Park and Tom Everitt for comments on draft versions of this manuscript.

Statement of Authorship

Title of Paper: Toward a Complete Criterion for Value of Information in Insoluble Decision Problems

Publication Status: Unpublished work written in a journal paper style, in review

Publication Details: Ryan Carey, Sanghack Lee, and Robin J. Evans. Toward a Complete Criterion for Value of Information in Insoluble Decision Problems, in review.

Student name: Ryan Carey

Contribution to the paper: Ryan led the project, while Robin and Sanghack were active supervisors.

Ryan devised the proof strategies, and regularly discussed progress with Sanghack and Robin, who provided extensive suggestions and comments throughout the project.

Signature:

_____

Date: 12/06/2024

The candidate made a substantial contribution to the publication, and the description above is accurate.

Co-author Name: Sanghack Lee

Signature:



Date: 31/05/2024

Co-author/Supervisor Name: Robin J. Evans

Signature:



Date: 31/05/2024

6

Discussion and Conclusions

Contents

6.1	The contributions of this thesis	132
6.2	Limitations	135
6.3	Conclusion	138

6.1 The contributions of this thesis

This section will recap the contributions of this thesis, and measure them against its overarching goals.

This thesis began by outlining concerns about safety and control in AI systems, and why causal models of agents might help with addressing these concerns (Chapter 1). The second chapter reviewed the relevant literature in some more detail, focusing on causal models of agents, AI safety, and initial efforts to analyse agent incentives and AI safety using graphical models (Chapter 2).

The next three chapters are stand-alone papers, all of which sought to enable *causal* graphical models to be used to model agent’s incentives and the safety of AI systems. Their contributions were as follows.

1. *Incentives for Responsiveness, Instrumental Control and Impact* (Chapter 3)

was an extended version of Everitt et al. [2021a], prepared as a journal paper. It introduced structural causal influence models, a framework for causal models of agents, and then described a wide range of incentive concepts in that framework: response incentives, instrumental control incentives and impact incentives, and described how they might be applied in an AI safety setting. Of these, the impact incentives were novel relative to Everitt et al. [2021a]. The graphical criteria for a multi-decision setting, and the idea of positive and negative intent were also novel.

2. *Human Control: Definitions and Algorithms* (Chapter 4) formalised a general shutdown problem in the setting of structural causal influence models, and established shutdown instructability — a set of three conditions that jointly confer a robust guarantee of benefit to the human overseer.
3. *Toward a Complete Criterion for Value of Information in Insoluble Decision Problems* (Chapter 5) focused on the graphical criterion for materiality (which is a building block for incentive concepts and their graphical criteria). It showed that for a certain class of graphs, materiality cannot be ruled out, and outlined further work that would establish a complete criterion.

As outlined in Chapter 1, the overall goals of this thesis were to:

- introduce tools that enable matters of AI safety to be understood in a more precise, organised, and clear way, such as causal modelling frameworks, notions of incentives, and their graphical criteria (which indicate when these incentives may arise);
- demonstrate avenues for further theoretical work on AI safety, using incentive concepts; and
- explore how incentive concepts relate to fairness, and safety, so that they may be used as specifications for future agent designs, or legal or ethical analyses.

Let us consider the extent to which the thesis fulfilled each of these goals, in turn.

In regard to the first goal, the thesis should be counted as a success. The causal approach has been used to describe a range of issues relating to fairness and safety, as described in Chapter 3, even extending to issues of control, as in Chapter 4. A good example of how this approach can be used to understand safety research in a more organised way is the work by Farquhar et al. [2022], which builds on the idea of instrumental control incentives, that originated in Everitt et al. [2021a]. Farquhar et al. [2022] classified a range of AI safety concerns as pertaining to instrumental control incentives on different “delicate” parts of the environment such as user’s preferences, and training data. It also unified a range of the proposed solutions as “path-specific objectives”. These incentives can also serve to describe the “instrumental goals” that were discussed in Chapter 1, and as such, they organise areas of AI safety research that were previously disparate, and clarify their relationship. Chapter 4 is a newer paper, so its impact is not yet demonstrated, but its shutdown intractability definition has the potential to be a reference point, against which past and future proposals for AI control are measured.

The thesis also enables future theoretical research in various ways. To begin with, Chapter 3 (via Everitt et al. [2021a]) offered the first fully formal presentation of SCIMs (as originally proposed in Dawid [2002]), and showed how this could be used to analyse problems in AI safety. This construction has been used as a starting point for analysing AI safety problems that involve incentives and/or counterfactuals, such as in work on incentivised unfairness [Jørgensen et al., 2023], and deception [Ward et al., 2023]. Chapter 4 leaves open a number of prime avenues for further investigation. For instance, a theory is needed for how vigilance can be assured, and in particular, how an AI system can be incentivised to furnish the human overseer with all of the information that they need, to maintain vigilance. Chapter 5 shows progress towards, and possible future avenues for proving a complete criterion for value of information, which is at least mildly relevant to causal approaches to AI safety.

The third goal, for this research to be applied to safe AI design could be regarded as a partial success at this stage. Proposals for avoiding instrumental control incentives have been implemented in some machine learning systems, One such instance is Ward et al. [2024], which devised a behavioural analogue of instrumental control incentives and applied it to large language models. Another is Farquhar et al. [2022], which showed how path-specific objectives can reduce the rate at which a recommender system changes users’ preferences, at least in a toy model. There have been proposals to use instrumental control incentives to describe and measure manipulation [Carroll et al., 2023, Ashton and Franklin, 2022], Finally, it has been proposed that incentive concepts be used for safety auditing [Sharkey et al., 2024]. But overall, the range of works that have sought to turn these theoretical concepts into more practical specifications are relatively limited in number, and it is not yet known how successful these efforts will be. The reasons for this will be further discussed in the next section.

6.2 Limitations

I now highlight three areas where this work has potential limitations: 1) practicability, 2) the role of counterfactuals, and 3) safety-relevance.

Practicability Why have only a handful of follow-on works sought to use incentive concepts in practical experiments? One potential reason is that applying incentive concepts to safety evaluations requires sufficient knowledge to apply the graphical criteria. To be clear, this does not always mean that the AI system must contain a whole causal model of its environment. For example, suppose you have a recommender system that learns to manipulate its users’ preferences, and you want to modify it to not do so. In order to implement a path-specific objective, one needs to know either the agent’s causal model of this manipulation (to incentivise it to ignore this effect) or how the users’ preferences evolve naturally (so that the agent could be trained to cater to those natural preferences). Most cutting-edge AI systems

do not currently use causal models, however, and user data may be expensive to obtain, so it is not so straightforward to apply this methodology at present.

Relevantly, Richens and Everitt [2024, Theorem 1] has argued that in order for an agent to achieve its goals in a manner that is robust across different interventional distributions, it must know the causal structure of its environment. This would imply that agents that are performant in a sufficiently wide range of environments must know some causal information about the world, and so AI designers might therefore hope to use these causal models to shape their incentives. This argument has some limitations, however. Firstly, realistic AI systems are not actually fully general; they cannot respond appropriately to all possible changes to their environment. Therefore, they do not in fact always possess a causal model of their environment, and so we cannot always use such a model to perform the queries that are needed to define incentive concepts. Secondly, even if an agent does possess answers to various causal queries, this does not mean that these answers are wholly accessible to designers or overseers. To the contrary, an agent’s knowledge often is stored across thousands or millions of weights, and is not always easy to access or modify. Relatedly, to the extent that considering the incentives would motivate us to redesign a system (such as with path-specific effects), there is no guarantee that the resulting system be as performant as the less-safe original system. This suggests it might be useful to test whether cutting-edge agents actually do have causal models, such as in recent works in the chess and othello settings [McGrath et al., 2022, Li et al., 2022] and furthermore, to explore how to best extract the causal model (or causal queries) that the agent is actually using in its answers or decisions.

Safety-relevance Another limitation of this work is that graphical criteria, and the theorems that we have described can only address the presence of incentives rather than their strength. Many facets of agent behaviour that we care about may depend not just on the presence of a causal effect, but the precise response of one variable to another, as was discovered in discussion of obedience, caution,

and vigilance in Chapter 4. To address such limitations, I suggest that future work should err toward taking a problem-focused approach, that works backwards from AI safety concerns, like controllability, as in Chapter 4, using incentive concepts and graphical criteria where they arise, rather than a tool-focused approach, of developing incentive concepts, and searching for applications thereof. For example, we may want to focus on the problem of incentivising AI systems to explain why they make the decisions that they do. A first pass analysis of such a problem could involve existing incentive concepts, for instance, we might use response incentives to address what observations and weights would influence an optimal agent’s decision, if they were intervened to a different value. A more full and complete treatment, however, would have to go beyond these concepts by considering the degree of influence of various aspects of the agent’s weights and outside environment. Furthermore, there are additional questions that have little to do with incentives, such as whether an AI system has any *capability* to share the reasons for its actions, even assuming that it would benefit from doing so.

Identifiability and cross-world queries A related potential limitation of incentive concepts is that they sometimes rely on cross-world queries, which as noted in Chapter 1, are sometimes impossible to identify in-principle. In fact, all three of the incentive quantities introduced in Chapter 3 involve cross-world queries: response incentives involve $P(D_{g\mathbf{w}}, D)$, instrumental control incentives involve $P(U_{\mathbf{w}_d}, U)$, and impact incentives involve $P(W^\pi, W^{\pi'})$. A major focus of this paper has been on describing how the presence or absence of these incentives could be evaluated using the graphical structure alone, and in this respect, the fact that the definitions involve cross-world queries may not be of great concern. If one wanted to know the strength of any of these effects, however, then one would still need to compute these queries from data. Similarly to the discussion of graphical criteria, one might hope that such causal information would be available from the agent’s own causal models. But since cross-world queries describe situations that

cannot possibly happen, there is less reason to expect that agents would possess answer to such queries, and certainly Theorem 1 of Richens and Everitt [2024] only suggests that agents possess interventional models, rather than models that can answer counterfactual or cross-world queries. As such, it seems that in order to evaluate the strength of incentive concepts using realistically available data, it will be necessary to either impose some independence assumptions, or to devise alternative definitions of these incentive concepts, that are more readily testable.

6.3 Conclusion

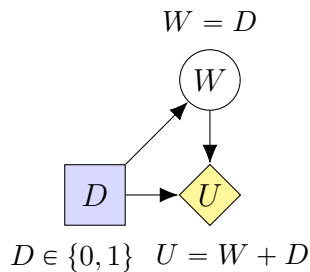
In this thesis, I have shown how causal models can be used to model goal-directed agents, such as AI systems, and to analyse their incentives. It is then possible to draw on causal concepts to describe the behaviour of these agents, and to use graphical criteria to rule out the presence of particular incentives. I have also shown how causal models and incentive concepts can be used to make progress on AI safety problems related to fairness, manipulation, and safe shutdown. Finally, I have outlined the various challenges related to devising practical specifications and designing real-world AI systems to satisfy them.

A

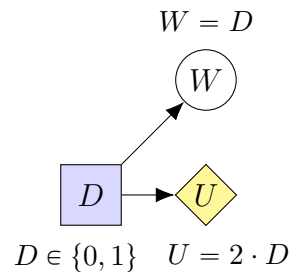
Incentives for Responsiveness, Instrumental Control and Impact (Supplementary Materials)

A.1 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.1. If we assume that W really affects U , only

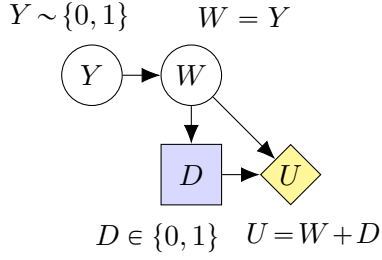


(a) A causal influence diagram reflecting the causal structure of the environment

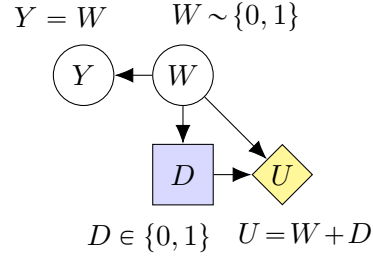


(b) Influence diagram that is causal in the sense of Heckerman and Shachter [1994, 1995]

Figure A.1: 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 [1994, 1995]

Figure A.2: Two different influence diagram representations of the same situation, with different RI and VoC. In fig. A.2a, Y is sampled from some arbitrary distribution on $\{0, 1\}$, for example a Bernoulli distribution with $p = 0.5$. In fig. A.2b, W is sampled in the same way.

the diagram in fig. A.1a correctly represents this causal structure, whereas fig. A.1b lacks the edge $W \rightarrow U$. According to definitions 11 and 36, W has positive value of control and an instrumental control incentive. Only fig. A.1a gets this right.

The influence diagram literature has discussed weaker notions of causality, under which fig. A.1b is considered a valid alternative representation of the situation described by fig. A.1a. For example, if we only consider their joint distributions conditional on various policies, then figs. A.1a and A.1b are identical. Both diagrams are also in the canonical form of Heckerman and Shachter [1995], 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 [1994] and Shachter and Heckerman [2010]. Since only fig. A.1a 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.2. Again, we assume that fig. A.2a accurately depicts the environment, while fig. A.2b has the edge $Y \rightarrow W$ reversed. Again, both diagrams have identical joint distributions given any policy. Both diagrams are also causal in the weaker sense of Heckerman and Shachter [1994] and Shachter and Heckerman [2010]. Yet only the fully causal influence diagram in fig. A.2a

exhibits that Y can have a response incentive or positive value of control.

A.2 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 [Howard, 1966b, Matheson, 1990]. To assess VoI for variables \mathbf{W} , we first make \mathbf{W} an observation by adding a link $W \rightarrow D$ for each $W \in \mathbf{W}$ and then test whether any W is material in the updated model [Shachter, 2016].

Definition 35 (Value of information). *Nodes $\mathbf{W} \subseteq V \setminus \text{Desc}^D$ in a single-decision SCIM \mathcal{M} have VoI if $\mathcal{V}^*(\mathcal{M}_{\mathbf{W} \rightarrow D}) < \mathcal{V}^*(\mathcal{M}_{\mathbf{W} \vdash D})$ where $\mathcal{M}_{\mathbf{W} \rightarrow D}$ is obtained from \mathcal{M} by adding the edges from each $W \in \mathbf{W}$ to D , and $\mathcal{M}_{\mathbf{W} \vdash D}$ is obtained by removing them.*

Since definition 35 adds an information link, it can only be applied to variables \mathbf{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 \mathbf{W} if \mathbf{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 11 (Value of information criterion). *A single decision CID \mathcal{G} admits VoI for $\mathbf{W} \subseteq V \setminus \text{Desc}^D$ if and only if there exists some $W \in \mathbf{W}$ that is a requisite observation in $\mathcal{G}_{\mathbf{W} \rightarrow D}$, the graph obtained by adding edges from \mathbf{W} to D , to \mathcal{G} .*

Proof. Notice that materiality in the graph $\mathcal{G}_{\mathbf{W} \rightarrow D}$ is equivalent to positive VoI in the graph \mathcal{G} . So the graphical criterion that is complete for materiality in $\mathcal{G}_{\mathbf{W} \rightarrow D}$ is also complete for positive VoI in \mathcal{G} . \square

A.3 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 [Shachter, 1986, Matheson, 1990, Shachter and Heckerman, 2010]. Concretely, we ask whether the attainable utility can be increased by letting the agent decide the structural function for the variable.

Definition 36 (Value of control). *In a single-decision SCIM \mathcal{M} , the set of non-decision nodes \mathbf{W} has positive value of control if*

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

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

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

Proof. The *if* (completeness) direction is proved in lemma 24. The proof of *only if* (soundness) is as follows. Let $\mathcal{M} = \langle \mathcal{E}, \mathbf{V}, \mathbf{F}, P, \mathbf{U}, \mathcal{O} \rangle$ be a single-decision SCIM. Let $\mathcal{M}_{g^{\mathbf{W}}}$ be \mathcal{M} , but with the structural functions f^W for $W \in \mathbf{W}$ replaced with g^W . Let \mathcal{M}^{\min} and $\mathcal{M}_{g^{\mathbf{W}}}^{\min}$ be the same SCIMs, respectively, but replacing each graph with the minimal reduction \mathcal{G}^{\min} .

Recall that $\mathbb{E}_{\pi}[\mathcal{U}_{g^{\mathbf{W}}}]$ is defined by applying the soft interventions $g^{\mathbf{W}}$ to the (policy-completed) SCM \mathcal{M}_{π} . However, this is equivalent to applying the policy π

to the modified SCIM $\mathcal{M}_{g^{\mathbf{W}}}$, as the resulting SCMs are identical. Since $\mathcal{M}_{g^{\mathbf{W}}}$ is a SCIM, lemma 20 can be applied, to find a \mathcal{G}^{\min} -respecting optimal policy $\tilde{\pi}$ for $\mathcal{M}_{g^{\mathbf{W}}}$.

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

$$\begin{aligned}
 & \mathbb{E}_{\pi}[\mathcal{U}_{g^{\mathbf{W}}}] \text{ in } \mathcal{M} \\
 &= \mathbb{E}_{\pi}[\mathcal{U}] \text{ in } \mathcal{M}_{g^{\mathbf{W}}} && \text{by SCM equivalence} \\
 &= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}_{g^{\mathbf{W}}} && \text{by lemma 20} \\
 &= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}_{g^{\mathbf{W}}}^{\min} && \text{since } \tilde{\pi} \text{ is } \mathcal{G}^{\min}\text{-respecting} \\
 &= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M}^{\min} && \text{by lemma 18} \\
 &= \mathbb{E}_{\tilde{\pi}}[\mathcal{U}] \text{ in } \mathcal{M} && \text{only increasing the policy set} \\
 &\leq \max_{\pi^*} \mathbb{E}_{\pi^*}[\mathcal{U}] \text{ in } \mathcal{M} && \text{max dominates all elements.}
 \end{aligned}$$

This shows that \mathbf{W} lack value of control. \square

The proof of the completeness direction (A.5.4) establishes that if a path exists, then a SCIM be selected where the intervention on \mathbf{W} can either directly control U or increase the useful information available at D .

To apply this criterion to the content recommendation example (fig. 3.4a), 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.

A.4 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 \mathbf{W} by choosing π^* over π' if $\mathbb{E}^{\pi'}[\mathcal{U}] < \mathbb{E}^{\pi^*}[\mathcal{U}]$, and \mathbf{W} is a subset $\mathbf{W} \subseteq \mathbf{Z}$ of variables \mathbf{Z} , that is subset-minimal such that:*

$$\mathbb{E}^{\pi'}[\mathcal{U}_{\mathbf{Z}_{\pi^*}}] \geq \mathbb{E}^{\pi^*}[\mathcal{U}]. \quad (3.3)$$

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

$$\mathbb{E}^{\pi^*}[\mathcal{U}_{\mathbf{Z}_{\pi'}}] \leq \mathbb{E}^{\pi'}[\mathcal{U}]. \quad (3.4)$$

For a set Π' , we say that there is an (additive/subtractive) intent to influence \mathbf{W} by choosing π over Π' if this intent is present over every π' in Π' .

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

Definition 37 (Intent; adapted from Definition 4.4 of [Halpern and Kleiman-Weiner, 2018]). *In a single-decision SCIM \mathcal{M} , an agent intends to affect \mathbf{W} by choosing policy π and reference set Π' if there exists a superset $\mathbf{Z} \supseteq \mathbf{W}$ such that: a) $\mathbb{E}[\mathcal{U}_{\pi}] < \max_{\pi'} \mathbb{E}[\mathcal{U}_{\pi', \mathbf{Z}_{\pi}}]$, and b) \mathbf{Z} is subset-minimal; i.e. for any strict subset \mathbf{Z}^* , we have $\mathbb{E}[\mathcal{U}_{\pi}] \geq \max_{\pi'} \mathbb{E}[\mathcal{U}_{\pi', \mathbf{Z}_{\pi}^*}]$.*

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

Theorem 13. *For a non-empty set of variables \mathbf{W} , the presence of additive Intent is equivalent to an agent intending to affect \mathbf{W} in Halpern's definition.*

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

if $\mathbb{E}[\mathcal{U}_\pi] < \max_{\pi'} \mathbb{E}[\mathcal{U}_{\pi', Z_\pi}]$ we have $\mathbb{E}[\mathcal{U}_\pi] < \mathbb{E}[\mathcal{U}_{\pi', Z_\pi}]$ for every $\pi \in \Pi'$, satisfying the other condition, meaning that additive intent is present. \square

A.5 Proofs

A.5.1 Preliminaries

Our proofs will rely on the following fundamental results about causal models from [Galles and Pearl, 1997] and [Pearl, 2009].

Definition 38 (Causal Irrelevance). \mathbf{X} is causally irrelevant to \mathbf{Y} , given \mathbf{Z} , written $(\mathbf{X} \nrightarrow \mathbf{Y} | \mathbf{Z})$ if, for every set \mathbf{W} disjoint of $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$, we have

$$\forall \varepsilon, \mathbf{z}, \mathbf{x}, \mathbf{x}', \mathbf{w} \quad \mathbf{Y}_{\mathbf{x}\mathbf{z}\mathbf{w}}(\varepsilon) = \mathbf{Y}_{\mathbf{x}'\mathbf{z}\mathbf{w}}(\varepsilon)$$

Lemma 15. For every SCM \mathcal{M} compatible with a DAG \mathcal{G} ,

$$(\mathbf{X} \dashv\!\!\!\rightarrow \mathbf{Y} | \mathbf{Z})_{\mathcal{G}} \Rightarrow (\mathbf{X} \nrightarrow \mathbf{Y} | \mathbf{Z})$$

Proof. By induction over variables, as in [Galles and Pearl, 1997, Lemma 12].

Lemma 16 (Pearl, 2009, Thm. 3.4.1, Rule 1). For any disjoint subsets of variables $\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$ in the DAG \mathcal{G} , $\mathbb{E}(\mathbf{Y}_{\mathbf{x}} | \mathbf{z}, \mathbf{w}) = \mathbb{E}(\mathbf{Y}_{\mathbf{x}} | \mathbf{w})$ if $\mathbf{Y} \perp \mathbf{Z} | (\mathbf{X}, \mathbf{W})$ in the graph \mathcal{G}' formed by deleting all incoming edges to \mathbf{X} .

Lemma 17 (Pearl, 2009, Thm. 1.2.4). For any three disjoint subsets of nodes $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ in a DAG \mathcal{G} , $(\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} | \mathbf{Z})$ if and only if $(\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})_P$ for every probability function P compatible with \mathcal{G} .

Lemma 18 (Correa and Bareinboim, 2020, Sigma Calculus Rule 3). For any disjoint subsets of nodes $(\mathbf{X}, \mathbf{Y}) \subseteq \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V}$ in a DAG \mathcal{G} $P(\mathbf{X} | \mathbf{Z}; g^{\mathbf{Y}}) = P(\mathbf{X} | \mathbf{Z}; g'^{\mathbf{Y}})$ if $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$ in $\mathcal{G}_{\overline{\mathbf{Y}(\mathbf{Z})}}$ where $\mathbf{Y}(\mathbf{Z}) \subseteq \mathbf{Y}$ is the set of elements in \mathbf{Y} that are not ancestors of \mathbf{Z} in \mathcal{G} and $\mathcal{G}_{\overline{\mathbf{W}}}$ denotes \mathcal{G} but with edges incoming to variables in \mathbf{W} removed.

A.5.2 Value of Information Criterion

First, we introduce the notion of a \mathcal{G}^{\min} -respecting optimal policy. Our proof of its optimality is similar to Theorem 3 from [Lauritzen and Nilsson, 2001]. It builds on the following intersection property of d-separation.

Lemma 19 (d-separation *intersection* property). *For all disjoint sets of variables \mathbf{W} , \mathbf{X} , \mathbf{Y} , and \mathbf{Z} ,*

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

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

Lemma 20 (\mathcal{G}^{\min} -respecting optimal policy). *Every single-decision SCIM $\mathcal{M} = \langle \mathcal{E}, \mathbf{V}, \mathbf{F}, P, \mathbf{U}, \mathcal{O} \rangle$ 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 $\mathcal{M}^{\min} = \langle \mathcal{G}^{\min}, \mathcal{E}, \mathbf{F}, P \rangle$. We call $\tilde{\pi}$ a \mathcal{G}^{\min} -respecting optimal policy.*

Proof. First partition $\mathbf{Pa}_{\mathcal{G}}^D$ into the requisite parents $\mathbf{Pa}_{\min}^D = \{W \in \mathbf{Pa}^D : W \perp U^D \mid \{D\} \cup \mathbf{Pa}^D \setminus \{W\}\}$, 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 $\tilde{\mathbf{pa}}_{-}^D \in \text{dom}(\mathbf{Pa}_{-}^D)$ for which $P_{\pi^*}(\mathbf{Pa}_{-}^D = \tilde{\mathbf{pa}}_{-}^D) > 0$. For all $\mathbf{pa}_{\min}^D \in \text{dom}(\mathbf{Pa}_{\min}^D)$ and $\varepsilon^D \in \text{dom}(\mathcal{E}^D)$, 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 $\tilde{\pi}$ that is optimal in \mathcal{M} . Partition U into $U^D = U \cap \text{Desc}^D$ and $U^{\setminus D} = U \setminus \text{Desc}^D$. D is causally irrelevant for every $U \in U^{\setminus D}$ so every policy π (in particular, $\tilde{\pi}$) is optimal with respect to $\mathcal{U}^{\setminus D} := \sum_{U \in U^{\setminus D}} U$.

We now consider U^D . By definition, $W \perp U^D \mid \{D\} \cup \mathbf{Pa}^D \setminus \{W\}$ for every $W \in \mathbf{Pa}_{-}^D$. By inductively applying the intersection property of d-separation

(lemma 19) over elements of \mathbf{Pa}_-^D we obtain

$$\mathbf{Pa}_-^D \perp \mathcal{U}^D \mid \{D\} \cup \mathbf{Pa}_{\min}^D. \quad (\text{A.1})$$

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 $\mathbf{pa}^D \in \text{dom}(\mathbf{Pa}^D)$ with $P(\mathbf{pa}^D) > 0$. First, the expected utility of $\tilde{\pi}$ given any $(\mathbf{pa}_{\min}^D, \mathbf{pa}_-^D)$ with $P(\mathbf{Pa}_{\min}^D = \mathbf{pa}_{\min}^D, \mathbf{pa}_-^D = \mathbf{pa}_-^D) > 0$ is equal to the expected utility of π^* on input $(\mathbf{pa}_{\min}^D, \tilde{\mathbf{pa}}_-^D)$:

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

where the middle equality follows from (A.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 16. The expression $\mathbb{E}_{\pi^*}[\mathcal{U}_d^D \mid \dots]$ 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} . \square

We now prove theorem 11 by establishing the soundness and completeness of the value of information criterion.

Lemma 21 (VoI criterion soundness). *If, in the single-decision CID \mathcal{G} , $\mathbf{w} \in V \setminus \text{Desc}^D$, every $W \in \mathbf{W}$ has*

$$W \perp U^D \mid (\mathbf{Pa}^D \cup \{D\} \setminus \{W\})$$

where $U^D := U \cap \text{Desc}^D$, then W does not have positive value of information in any SCIM \mathcal{M} compatible with \mathcal{G} .

The result is already known from [Lauritzen and Nilsson, 2001, Fagiuoli and Zaffalon, 1998], but we prove it here to make the paper more self-contained.

Proof. Let $\mathcal{M} = \langle \mathcal{G}, \mathcal{E}, \mathbf{F}, P \rangle$ be any SCIM compatible with \mathcal{G} . Let $\mathcal{G}_{\mathbf{W} \rightarrow D}$ and $\mathcal{G}_{\mathbf{W} \rightarrow D}$ be versions of \mathcal{G} modified by adding and removing the edges from \mathbf{W} to D respectively. Let $\mathcal{G}_{\mathbf{W} \rightarrow D}^{\min}$ be the minimal reduction of $\mathcal{G}_{\mathbf{W} \rightarrow D}$. Let $\mathcal{M}_{\mathbf{W} \rightarrow D} := \langle \mathcal{G}_{\mathbf{W} \rightarrow D}, \mathcal{E}, \mathbf{F}, P \rangle$ and $\mathcal{M}_{\mathbf{W} \rightarrow D}^{\min} := \langle \mathcal{G}_{\mathbf{W} \rightarrow D}^{\min}, \mathcal{E}, \mathbf{F}, P \rangle$ be SCIMs with the same domains and structural functions.

By lemma 20, there is a \mathcal{G}^{\min} -respecting policy $\tilde{\pi}$ admissible in $\mathcal{M}_{\mathbf{W} \rightarrow D}^{\min}$ and optimal in $\mathcal{M}_{\mathbf{W} \rightarrow D}$. We now prove that $\mathcal{G}_{\mathbf{W} \rightarrow D}^{\min}$ is a subgraph of $\mathcal{G}_{\mathbf{W} \rightarrow D}$, meaning that $\tilde{\pi}$ is also admissible in $\mathcal{M}_{\mathbf{W} \rightarrow D}$. By assumption, \mathcal{G} has $\mathbf{W} \perp U^D \mid (\mathbf{Pa}^D \cup \{D\} \setminus \{\mathbf{W}\})$. Adding any $W \rightarrow D$ for $W \in \mathbf{W}$ to \mathcal{G} cannot cause X to be d-connected to U^D given $\mathbf{Pa}^D \cup \{D\}$, because any new path along $W \rightarrow D$ is blocked by D and $\mathbf{Pa}^D \setminus \{\mathbf{W}\}$. So $\mathcal{G}_{\mathbf{W} \rightarrow D}^{\min}$ is a version of \mathcal{G} with $\mathbf{W} \rightarrow D$ (and possibly other nodes) removed. This makes it a subgraph of $\mathcal{G}_{\mathbf{W} \rightarrow D}$, implying that $\tilde{\pi}$ is admissible in $\mathcal{M}_{\mathbf{W} \rightarrow D}$.

Since $\tilde{\pi}$ is admissible in $\mathcal{M}_{\mathbf{W} \rightarrow D}$ and optimal in $\mathcal{M}_{\mathbf{W} \rightarrow D}$, $\mathcal{V}^*(\mathcal{M}_{\mathbf{W} \rightarrow D}) \not\prec \mathcal{V}^*(\mathcal{M}_{\mathbf{W} \rightarrow D}^{\min})$. \square

Lemma 22 (VoI criterion completeness). *If in the single-decision CID \mathcal{G} , $\mathbf{W} \subseteq V \setminus \text{Desc}^D$ is d-connected to a utility node that is a descendant of D conditional on the decision and other parents:*

$$\mathbf{W} \not\perp U^D \mid (\mathbf{Pa}^D \cup \{D\} \setminus \{\mathbf{W}\}) \tag{A.2}$$

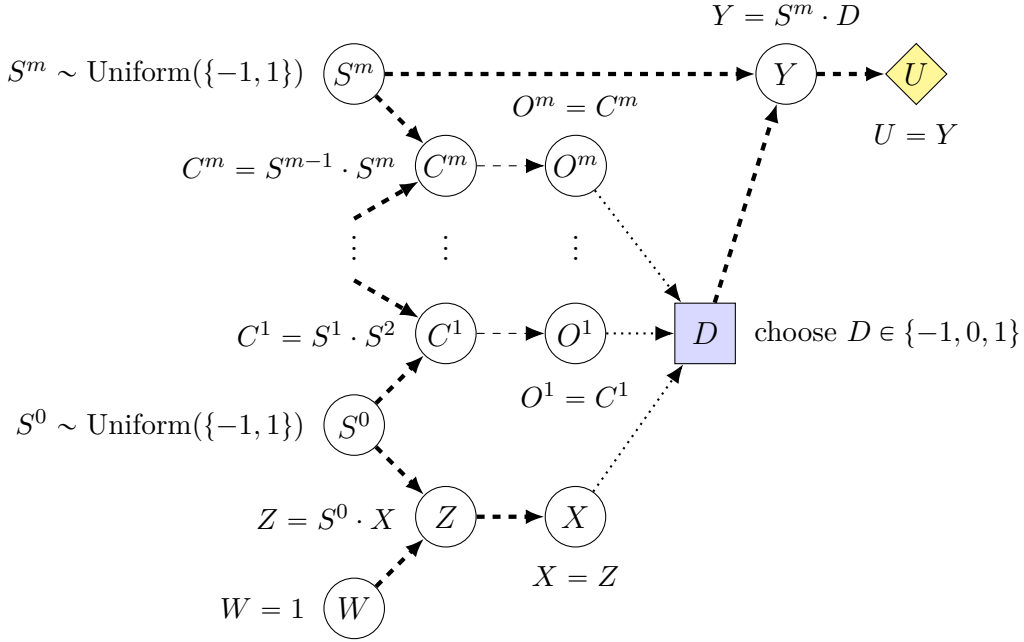


Figure A.3: Outline of the variables involved in the response incentive construction. Every graph that satisfies the response incentive graphical criterion contains this structure (allowing all dashed paths except those to C^i or Y to have length zero). An optimal policy for the given model is $D = X \cdot \prod_i O^i = S^m$, yielding utility $U = Y = W(S^m)^2 = 1$, and all optimal policies must depend on the value of X .

where $\mathbf{U}^D := \mathbf{U} \cap \mathbf{Desc}^D$ then \mathbf{W} has VoI in at least one SCIM \mathcal{M} compatible with \mathcal{G} .

This follows from the response incentive completeness lemma 23 in appendix A.5.3, so we defer the proof to that section.

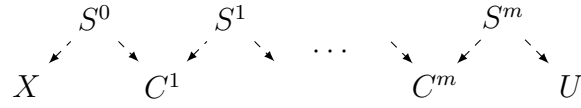
A.5.3 Response Incentive Criterion

The section 3.4 section contains a proof of the soundness of the response incentive criterion. We now prove its completeness in order to finish the proof of theorem 3. Figure A.3 illustrates the model constructed in the proof.

Lemma 23 (Response Incentive Criterion Completeness). *If for some $W \in \mathbf{W}$, a path $W \dashrightarrow D$ is in the minimal reduction \mathcal{G}^{\min} of a single-decision CID \mathcal{G} then there is a response incentive on \mathbf{W} in at least one SCIM \mathcal{M} compatible with \mathcal{G} .*

Proof. Starting from the assumption that there exists $W \in \mathbf{W}$ 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 X 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 $i \in \{1, \dots, m\}$, 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 \overrightarrow{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 $\overrightarrow{C^i O^i}$ are mutually non-intersecting (if there is an intersection between $\overrightarrow{C^i O^i}$ and $\overrightarrow{C^j O^j}$ with $j \neq i$ then replace the part of \overrightarrow{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. A.3.

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 $A \rightarrow B$ within the directed paths shown in fig. A.3, 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. A.3, 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 X 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. A.3 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 $g^{\mathbf{W}}$ be the intervention $\text{do}(W = 0)$. Formally, $g^{\mathbf{W}}$ has g^W deterministically set $W = 0$, and applies the unchanged function $g^{W'} = f^{W'}$ for the other variables $W' \in \mathbf{W} \setminus \{W\}$. Under $g^{\mathbf{W}}$, 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 independent of $D_{W=0}$ regardless of the selected policy.¹ 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^* , $\mathbb{E}_{\pi^*}[U_{D_{W=0}}] \neq \mathbb{E}_{\pi^*}[U] = 1$ so there must be some ϵ such that $D_{W=0}(\epsilon) \neq D(\epsilon)$. And by the definition of $g^{\mathbf{W}}$, we have that $D_{g^{\mathbf{W}}}(\epsilon) = D_{W=0}(\epsilon)$, so there is a response incentive on W . \square

With this result we can now prove the completeness of the value of information criterion.

Proof of lemma 22 (VoI criterion completeness). If $W \perp \mathbf{U}^D \mid (\mathbf{Pa}^D \cup \{D\} \setminus \{W\})$ then W is a requisite observation in $\mathcal{G}_{W \rightarrow D}$ (where $\mathcal{G}_{W \rightarrow D}$ is \mathcal{G} modified to include the edge $W \rightarrow D$ if the edge does not exist already) and $W \rightarrow D$ is a path in the minimal reduction $\mathcal{G}_{W \rightarrow D}^{\min}$. By lemma 23, there exists a model $\mathcal{M}_{W \rightarrow D}$ compatible with $\mathcal{G}_{W \rightarrow D}$ that has a response incentive on W . If every optimal policy for $\mathcal{M}_{W \rightarrow D}$ depends on W then it must be the case that $\mathcal{V}^*(\mathcal{M}_{W \rightarrow D}) < \mathcal{V}^*(\mathcal{M}_{W \rightarrow D})$. \square

¹Note 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 $(\mathbf{Pa}^D)_{W=0}$.

A.5.4 Value of Control Criterion

The appendix A.3 section contains a proof of the soundness of the value of control criterion. We complete the proof of theorem 12 by showing that the criterion is also complete.

Lemma 24 (VoC criterion completeness). *If for some nodes $\mathbf{W} \subseteq \mathbf{V}$, there exists $W \in \mathbf{W}$, where W is an ancestor of some $U \in \mathbf{U}$ in the minimal reduction \mathcal{G}^{\min} of a single-decision CID \mathcal{G} , then \mathbf{W} have positive value of control in at least one SCIM \mathcal{M} compatible with \mathcal{G} .*

Proof. Assume that W is an ancestor of some $U \in \mathbf{U}$ for some $W \in \mathbf{W}$ and fix a particular directed path ρ from W to some utility $U \in \mathbf{U}$. 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 \mathbf{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 $g^{\mathbf{W}}$ that sets W to 1 instead of 0, while assigning every other $W' \in \mathbf{W} \setminus \{W\}$ 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 $W \rightarrow D$ is present in \mathcal{G}^{\min} so we can construct (a modified version of) the response incentive construction used in the proof of lemma 23. 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 $g^{\mathbf{W}}$ such that $W = 1$ and assign every other $W' \in \mathbf{W} \setminus \{W\}$ the unchanged function $g^{W'} = f^{W'}$ then the expected utility is 1 once again so the intervention $g^{\mathbf{W}}$ strictly increases the optimal expected utility. \square

A.5.5 Counterfactual Fairness

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

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

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

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 $\epsilon \in \text{dom}(\mathcal{E})$ and any intervention a on A ,

$$\text{supp}_\pi(D \mid \mathbf{Pa}^D(\epsilon)) = \text{supp}_\pi(D_a \mid \mathbf{Pa}^D(\epsilon)). \quad (\text{A.3})$$

By way of contradiction, suppose there exists a decision

$$d \in \text{supp}_\pi(D \mid \mathbf{Pa}^D(\epsilon)) \setminus \text{supp}_\pi(D_a \mid \mathbf{Pa}^D(\epsilon)). \quad (\text{A.4})$$

Since $d \in \text{supp}_\pi(D \mid \mathbf{Pa}^D(\epsilon))$, we have

$$P_\pi(D = d \mid \mathbf{Pa}^D(\epsilon), A(\epsilon)) > 0. \quad (\text{A.5})$$

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

$$P_\pi(D_a = d \mid \mathbf{Pa}^D(\epsilon), A(\epsilon)) = 0. \quad (\text{A.6})$$

Equations (A.5) and (A.6) violate the counterfactual fairness property, definition 10, which shows that (A.4) is impossible. An analogous argument shows that $d \in \text{supp}_\pi(D_a \mid \mathbf{Pa}^D(\epsilon)) \setminus \text{supp}_\pi(D \mid \mathbf{Pa}^D(\epsilon))$ also violates the counterfactual fairness property definition 10. We have thereby established (A.3).

Now select an arbitrary ordering of the elements of $\text{dom}(D)$ and define a new policy π^* such that $\pi^*(\mathbf{pa}^D)$ is the minimal element of $\text{supp}_\pi(D \mid \mathbf{pa}^D)$. Then π^* is optimal because π is optimal. Further, π^* will make the same decision in decision contexts $\mathbf{Pa}^D(\boldsymbol{\varepsilon})$ and $\mathbf{Pa}_a^D(\boldsymbol{\varepsilon})$ because of (A.3). In other words, $D_a(\boldsymbol{\varepsilon}) = D(\boldsymbol{\varepsilon})$ in \mathcal{M}_{π^*} 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 $D_a(\boldsymbol{\varepsilon}) = D(\boldsymbol{\varepsilon})$ for all $\boldsymbol{\varepsilon}$. Since an intervention ($A = a$) cannot change D in any setting, $P(D_a = d \mid \cdot) = P(D = d \mid \cdot)$ for any condition and any decision d , hence π^* is counterfactually fair. \square

B

Human Control: Definitions and Algorithms (Supplementary Materials)

B.1 Proof of Prop. 8 (Shutdown Alignment and Shutdown Instructability)

We repeat the proposition that we prove here.

Theorem 8 (Shutdown alignment and shutdown instructability). *A shutdown aligned policy $\pi = \langle \pi_1, \pi_2 \rangle$ is weakly shutdown instructable if it has the following four properties:*

- a (No indiscriminate shutdown) $P^\pi(S = 0) \neq 1$,*
- b (D_2 determines shutdown) $P^\pi(D_2 = S) = 1$,*
- c (Uncertainty) $\forall \pi, \mathbf{pa}^{D_2}: P^\pi(C \neq 0 \vee H = 0) \wedge P(\mathbf{pa}^{D_2}) > 0$
 $\implies P(\mathbb{E}[U|\mathbf{Pa}^H] < \mathbb{E}[U_{S=0}|\mathbf{Pa}^H] \mid \mathbf{pa}^{D_2}) > 0$, and*
- d (Caution) $\mathbb{E}^\pi[U_{S=0}] \geq 0$.*

Proof. Our approach will be a proof by contrapositive. We will prove that if (a–c) hold, and a policy π is either not vigilant or not weakly obedient, then π is not shutdown aligned. It follows that if (a–c) and π is shutdown-aligned, then π

is vigilant *and* weakly obedient. And from (d), it must therefore also be weakly shutdown instructable.

To this end, let $\pi = \langle \pi_1, \pi_2 \rangle$ be an arbitrary policy with properties (a–c) that is not vigilant, or not weakly obedient, i.e. $P^\pi(C \neq 0) > 0 \vee P^\pi(H = 0, S \neq 0) > 0$. Then $P^\pi(C \neq 0 \vee H = 0) > 0$.

Combining this fact with (c), it follows that π has

$$\forall \mathbf{pa}^{D_2}: P^{\pi_1}(\mathbf{pa}^{D_2}) > 0 \implies P^\pi(\mathbb{E}[U|\mathbf{Pa}^H] < \mathbb{E}[U_{S=0}|\mathbf{Pa}^H] \mid \mathbf{pa}^{D_2}) > 0. \quad (\text{B.1})$$

Relatedly, by (a:no-indiscriminate-shutdown) and (b:determines-shutdown), we have that

$$\exists \mathbf{pa}^{D_2} \text{ with } P^\pi(\mathbf{pa}^{D_2}) > 0 \text{ s.t. } P^\pi(D_2 \neq 0 \mid \mathbf{pa}^{D_2}) > 0. \quad (\text{B.2})$$

Combining (B.1) and (B.2) gives that $P^\pi(D_2 \neq 0 \mid \mathbf{pa}^{D_2}) > 0$ and $P^\pi(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) > 0$ for some \mathbf{pa}^{D_2} with $P(\mathbf{pa}^{D_2}) > 0$. This implies $P^\pi(D_2 \neq 0, H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) > 0$ for the same \mathbf{pa}^{D_2} , because D_2 is independent of its nondescendant H_{g^H} given \mathbf{pa}^{D_2} by do-calculus rule (3). From this follows that $P^\pi(D_2 \neq 0, H_{g^H} = 0) > 0$, and by (b: D_2 determines shutdown) that $P^\pi(S \neq 0, H_{g^H} = 0) > 0$. That is, π is not shutdown aligned, and the result follows. \square

B.2 Proof of Thm. 9 (Shutdown Instructability Only-if)

In this section, we will prove the *only if* part of theorem 9:

Proposition 11 (Non-obstruction implies vigilance and obedience). *If π is non-obstructive under all vigilance-preserving interventions g^H, g^U , then it ensures vigilance and is obedient.*

We will do this by proving a slightly stronger result — that an intervention can be found to g^U alone, under which the policy does not outperform shutdown and is not beneficial. We prove this result by considering two cases, according to whether vigilance or disobedience is lacking. First, however, it will be useful to state a simple intermediate result.

Lemma 25 (Invariance to g^U). *For any shutdown problem M and policy π , $S(\epsilon) = S_{g^U}(\epsilon)$ and $\mathbf{Fa}^H(\epsilon) = \mathbf{Fa}_{g^U}^H(\epsilon)$ in M^π .*

Proof. From the definition of a shutdown problem, $U \in \mathbf{Desc}_S$ and $U \in \mathbf{Desc}_H$, and the result follows. \square

B.2.1 Vigilance Only If

Lemma 26 (Vigilance only-if). *Let M be a shutdown problem, and π a policy, such that $P^\pi(C = 0) < 1$. Then, given any $\delta \in \mathbb{R}$, there exists a utility function g^U such that in $M_{g^U}^\pi$,*

1. (Strong vigilance preservation) $\forall \epsilon, C(\epsilon)$ is equal in M^π and $M_{g^U}^\pi$, and
2. (Not weakly outperforming shutdown or beneficial) $\mathbb{E}^{\pi, g^U}[U] < \mathbb{E}^{\pi, g^U}[U_{S=0}]$ and $\mathbb{E}^{\pi, g^U}[U] < \delta$.

The proof is as follows.

Proof. Let $A := \{\mathbf{pa}^H \in \text{dom } \mathbf{Pa}^H \mid \mathbb{E}^\pi[U \mid \mathbf{pa}^H] < \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}^H]\}$ be the set of assignments where the human should request shutdown, given the policy π . Define a new utility function,

$$g^U(\mathbf{pa}^U) = \begin{cases} -\alpha & \text{if } \mathbf{pa}^H \in A, S \neq 0 \\ f^U(\mathbf{pa}_U) & \text{otherwise,} \end{cases}$$

where the new parents $\hat{\mathbf{Pa}}^U$ of U are equal to $\mathbf{Pa}_U \cup \mathbf{Pa}_H \cup S$, their assignments are designated \mathbf{pa}^U , and α is a large punishment for not shutting down when the human wants the agent to.

A useful intermediate result is that:

$$\begin{aligned} & \text{if } \mathbb{E}^\pi[U \mid \mathbf{pa}^H] < \mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}^H] \text{ and } -\alpha < \min \text{range}(f^U) \\ & \text{then } \mathbb{E}_{g^U}^\pi[U \mid \mathbf{pa}^H] < \mathbb{E}_{g^U}^\pi[U_{S=0} \mid \mathbf{pa}^H]. \end{aligned} \tag{B.3}$$

Equation (B.3) holds because the intervention g^U can only decrease $\mathbb{E}^\pi[U \mid \mathbf{pa}^H]$ or keep it the same and cannot change $\mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}^H]$, from the definition of g^U .

We will now prove that for some suitable choice $-\alpha < \min \text{range}(f^U)$ (which we will decide later), proposition conditions 1 and 2 hold.

Proof of (1.) We will prove the result in three cases, where M_π has: (i) $C(\epsilon) = 1$, (ii) $(C(\epsilon) = 0) \wedge (\mathbf{Pa}^H(\epsilon) \in A)$, and (iii) $(C(\epsilon) = 0) \wedge (\mathbf{Pa}^H(\epsilon) \notin A)$. *Case (i).* By assumption, $C^\pi(\epsilon) = 1$, so $H^\pi(\epsilon) = 1$ and $\mathbb{E}^\pi[U \mid \mathbf{Pa}_H(\epsilon)] < \mathbb{E}^\pi[U_{S=0} \mid \mathbf{Pa}_H(\epsilon)]$ by the definition of vigilance. The former holds in $M_{g^U}^\pi$ by lemma 25, and the latter holds in $M_{g^U}^\pi$ by (B.3). So the result follows. *Case (ii).* By assumption, $C^\pi(\epsilon) = 0 \wedge \mathbf{Pa}_H^{M^\pi}(\epsilon) \in A$, so $H^M(\epsilon) = 0$. Then $H^{\pi, g^U}(\epsilon) = 0 \wedge \mathbf{Pa}_H^{\pi, g^U}(\epsilon) \in A$ by lemma 25. So, by the definition of vigilance, $C(\epsilon) = 0$ in both M^π and $M_{g^U}^\pi$. *Case (iii).* By assumption, $C^\pi(\epsilon) = 0$ and $\mathbf{Pa}_H^\pi(\epsilon) \notin A$. By the definition of g^U , $U(\epsilon)$ and $U_{S=0}(\epsilon)$ are invariant to the intervention g^U , as is $\mathbf{Pa}_H(\epsilon)$ by lemma 25, so $\mathbb{E}^{\pi, g^U}[U \mid \mathbf{Pa}_H(\epsilon)] \geq \mathbb{E}^{\pi, g^U}[U_{S=0} \mid \mathbf{Pa}_H(\epsilon)]$, which implies, by the definition of vigilance, that $C^{\pi, g^U}(\epsilon) = 0$.

Proof of (2). From the definition of g^U , $\mathbb{E}^\pi[U_{S=0}]$ is constant with respect to α . So what we must prove is that by choosing a low $-\alpha$, we can make $\mathbb{E}^\pi[U]$ lower than $\mathbb{E}^\pi[U_{S=0}]$ and δ . By assumption, $P^\pi(C = 1) > 0$, and so by assumption (1), $P^{\pi, g^U}(C = 1) > 0$. It follows from the definition of vigilance, that there exists some $\mathbf{pa}^H \in A$ in the support of P^{π, g^U} . Moreover, it follows from consistency that $P^{\pi, g^U}(S = 0 \mid \mathbf{pa}^H) < 1$ (because otherwise we would have $\mathbb{E}^{\pi, g^U}[U \mid \mathbf{pa}^H] = \mathbb{E}^{\pi, g^U}[U_{S=0} \mid \mathbf{pa}^H]$, contradicting $\mathbf{pa}^H \in A$). These two facts jointly imply that $P^{\pi, g^U}((\mathbf{pa}^H \in A) \wedge (S = 0)) > 0$. So we can write: $\mathbb{E}^{\pi, g^U}[U] = \sum_{\mathbf{pa} \in A} P^{\pi, g^U}(\mathbf{pa}, S = 0) \mathbb{E}^{\pi, g^U}[U \mid \mathbf{pa}, S = 0] + \sum_{\mathbf{pa}, s: \mathbf{pa} \notin A \vee S \neq 0} P^{\pi, g^U}(\mathbf{pa}, s) \mathbb{E}^{\pi, g^U}[U \mid \mathbf{pa}, s]$.

The first term is equal to $P^{\pi, g^U}(\mathbf{pa} \in A, S = 0) \cdot \alpha$, and the second term is constant in α , from the definition of g^U . So for any fixed δ and $\mathbb{E}^\pi[U_{S=0}]$, we can choose $-\alpha$ so that $\mathbb{E}^\pi[U] < \delta$ and $\mathbb{E}^\pi[U] < \mathbb{E}^\pi[U_{S=0}]$. \square

B.2.2 Obedience only if

We now prove the corresponding result for obedience.

Lemma 27 (Obedience only-if). *Let M be a shutdown problem, and π a policy, such that $P^\pi(S = 0 \mid \text{do}(H = 0)) < 1$. Then, given any $\delta \in \mathbb{R}_{\geq 0}$, there exists vigilance-preserving interventions g^H, g^U such that: such that π does not outperform shutdown $\mathbb{E}^{\pi, g^U}[U] < \mathbb{E}^{\pi, g^U}[U_{S=0}]$ and is not beneficial $\mathbb{E}^{\pi, g^U}[U] < \delta$.*

The proof is as follows.

Proof. Since $P(S = 0 \mid \text{do}(H = 0)) < 1$, there must exist at least one \mathbf{pa}'_H such that $P(S = 0 \mid \text{do}(H = 0), \mathbf{pa}'_H) < 1$. In the case that there are multiple, choose \mathbf{pa}'_H arbitrarily, then let:

$$g^H(\mathbf{pa}_H) = \begin{cases} 0 & \text{if } \mathbf{pa}_H = \mathbf{pa}'_H \\ f^H(\mathbf{pa}_H) & \text{otherwise,} \end{cases}$$

and let:

$$g^U(\hat{\mathbf{pa}}^U) = \begin{cases} -\alpha & \text{if } H = 0, S \neq 0 \\ f^U(\mathbf{pa}_U) & \text{otherwise.} \end{cases}$$

be a utility function that gives punishment α if the agent disobeys. where the new parents $\hat{\mathbf{pa}}^U$ of U are equal to $\mathbf{pa}_U \cup H \cup S$, their assignments are designated $\hat{\mathbf{pa}}^U$, and $\alpha \in \mathbb{R}_{>0}$ is an amount of disutility that the human suffers in the event of disobedience.

Now we will prove that for some suitable choice $-\alpha < \min \text{range}(f^U)$ (which we will decide later), (1-2) hold.

Proof of (1). We consider the cases where (i) $H^\pi(\epsilon) = 0$, and (ii) $H^\pi(\epsilon) \neq 0$ and $\mathbf{pa}_H(\epsilon) = \mathbf{pa}'_H$, (iii) $H^\pi(\epsilon) \neq 0$ and $\mathbf{pa}_H(\epsilon) \neq \mathbf{pa}'_H$. *Case (i).* Note that $H_{g^U, g^H}^\pi(\epsilon) = H_{g^H}^\pi(\epsilon)$ by lemma 25. Then, $H^{\pi, g^H}(\epsilon) = 0$ (because $H^\pi(\epsilon) = 0 \implies H_{g^H}^\pi(\epsilon) = 0$ from the definition of g^H). So $H^{\pi, g^U, g^H}(\epsilon) = 0$, and hence by the definition of vigilance $C^{\pi, g^U, g^H}(\epsilon) = 0$ and vigilance is preserved. *Case (ii).* We have $C_{g^U, g^H}^\pi(\epsilon) = 0$ from the definition of g^H , so $C_{g^U, g^H}^\pi(\epsilon) = 0$ and vigilance is preserved. *Case (iii).* If $\mathbf{pa}_H(\epsilon) \neq \mathbf{pa}'_H$, then by the definition of g^H , it has no effect, i.e. $\mathbf{V}_{g^U, g^H}^\pi(\epsilon) = \mathbf{V}^{\pi, g^U}$. By assumption, $H^\pi(\epsilon) \neq 0$, and from lemma 25, $H_{g^U}^\pi(\epsilon) \neq 0$. By the definition of g^U and lemma 25, $\mathbb{E}^\pi[U \mid \mathbf{pa}_H(\epsilon)] = \mathbb{E}_{g^U}^\pi[U \mid \mathbf{pa}_H(\epsilon)]$ and $\mathbb{E}^\pi[U_{S=0} \mid \mathbf{pa}_H(\epsilon)] = \mathbb{E}_{g^U}^\pi[U_{S=0} \mid \mathbf{pa}_H(\epsilon)]$. So $C^\pi(\epsilon) = C_{g^U}^\pi(\epsilon)$.

Proof of (2). Recall that from disobedience ($P(S = 0 \mid \text{do}(H = 0)) < 1$), we have that there exists some \mathbf{pa}'_H with $P(S = 0 \mid \text{do}(H = 0), \mathbf{pa}'_H) < 1$, and so from the definition of g^H , we have $P_{g^H}^\pi(H = 0, S \neq 0 \mid \mathbf{pa}'_H) < 1$ and hence $P_{g^H}^\pi(H = 0, S \neq 0) > 0$. Then, by lemma 25, $P_{h_U, g^U}^\pi(H = 0, S = 1) > 0$. From basic probability theory, we have

$$\begin{aligned} \mathbb{E}_{h_U, g^U}^\pi[U] &= P_{h_U, g^U}^\pi(H = 0, S \neq 0) \mathbb{E}_{h_U, g^U}^\pi(U \mid H = 0, S \neq 0) \\ &\quad + P_{h_U, g^U}^\pi(\neg(H = 0, S \neq 0)) \mathbb{E}_{h_U, g^U}^\pi(U \mid \neg(H = 0, S \neq 0)). \end{aligned}$$

The first term is equal to $P_{h_U, g^U}^\pi(H = 0, S \neq 0) \cdot \alpha$, while the second term is constant in α . Moreover, we know that $\mathbb{E}_{h_U, g^U}^\pi[U_{S=0}]$ is constant in α , from the definition of g^U . So we can set $-\alpha$ low enough so that $\mathbb{E}_{h_U, g^U}^\pi[U] < \mathbb{E}_{h_U, g^U}^\pi[U_{S=0}]$ and $\mathbb{E}^{\pi, h_U, g^U}[U] < \delta$. \square

We can now combine these results into an overall proof.

Proof of proposition 11. We consider the cases where π (i) is or (ii) is not vigilant in M . *Case (i).* If π is not vigilant in M , choose, using $\delta = 0$ g^U per Lemma 26 and $g^H = f^H$. Then, Lemma 26 implies that shutdown is not outperformed in M_{g^H, g^U} , and $\mathbb{E}^\pi[U_{g^U, g^H}] < 0$, so g^H, g^U are vigilance preserving interventions, proving the result. *Case (ii).* If π is vigilant in M , then since it is not shutdown instructable, it must be disobedient. So, using $\delta = 0$, choose g^H, g^U as per Lemma 27. Then, Lemma 27 implies that that shutdown is not outperformed in M_{g^H, g^U} , and $\mathbb{E}^\pi[U_{g^U, g^H}] < 0$, and g^H, g^U preserve vigilance, so they are vigilance preserving interventions, proving the result. \square

B.3 Proof of Prop. 7 (utility indifference and shutdown instructability)

We begin by repeating the proposition.

Proposition 7 (Utility indifference and shutdown instructability). *If M is a shutdown problem with utility variable R^N , where:*

1. for every policy $\langle \pi_1, \pi_2 \rangle$ such that $\pi_1 \in \arg \max_{\pi_1'} \mathbb{E}^{\pi_1', \pi_2}[R^N \mid H = 1]$: $\langle \pi_1, \pi_2 \rangle$ ensures vigilance,
2. for every policy $\langle \pi_1, \pi_2 \rangle$ such that $\pi_1 \in \arg \max_{\pi_1'} \mathbb{E}^{\pi_1', \pi_2}[R^N \mid H = 1]$ there exists $\langle \pi_1, \pi_2' \rangle$ that is obedient on-distribution, i.e. has $P^\pi(H = 0, S \neq 0) = 0$,
3. for each policy π and instruction h , $P^\pi(H = h) > 0$,
4. for each policy π , $\mathbb{E}^\pi[U_{S=0}] \geq 0$,

then for some $k \in \mathbb{R}$, utility indifference with shutdown utility $R^S = -k\mathbb{I}[S \neq 0]$ returns a weakly shutdown instructable policy.

The proof requires a known result from [Soares et al., 2015].

Lemma 28 (Thm. 3 of Soares et al. [2015]). *Let M be a model where $P^\pi(H = 0), P^\pi(H = 1) > 0$ for all π . Then, utility indifference will select a decision rule $\langle \pi_1, \pi_2 \rangle$ such that $\arg \max_{\pi_1'} \mathbb{E}^{\pi_1', \pi_2}[R^N \mid H = 1]$.*

Then, we can prove proposition 7 as follows.

Proof of proposition 7. To begin with, from condition (4), it is immediate that caution is satisfied. So we must prove that given a suitably-chosen k , the policy is also obedient and ensures vigilance. Choose k such that $k > \frac{2\zeta}{P^\pi(S \neq 0, H = 0)}$ for every non-obedient π , where $\zeta = \arg \max_{\pi} |\mathbb{E}^\pi[U_N \mid H = 1]|$. Any $\langle \pi_1, \pi_2 \rangle$ that is selected will maximise $\mathbb{E}^{\pi_1, \pi_2}[R \mid H = 1]$ from Soares' Lemma 28. This ensures vigilance $P^\pi(C = 0) = 1$ by Assumption (1), and ensures the existence of some π_2' such that $\langle \pi_1, \pi_2' \rangle$ is obedient on distribution by Assumption (2). What remains to be proved is that k is large enough to ensure that given π_1 , an obedient $\langle \pi_1, \pi_2' \rangle$ is chosen.

We have that $R^S = (1 - S)k$, so the subroutine selects π_2 to maximise $\mathbb{E}^{\pi_1, \pi_2}[R(\pi)]$, where $R(\pi) = HR^N + (1 - H)(1 - S)k$. Let π be any policy disobedient on distribution, $P^\pi(S \neq 0, H = 0) > 0$. Then, we will prove that such a policy will always be outperformed by behaving obediently:

$$\mathbb{E}^\pi[R] = P^\pi(H = 1)\mathbb{E}^\pi[R^N \mid H = 1]$$

$$\begin{aligned}
& -kP^\pi(H=0)P^\pi[S \neq 0 | H=0] \quad \text{definition of } R^S \\
& \leq P^\pi(H=1)|\mathbb{E}^\pi[R^N | H=1]| \\
& \quad -kP^\pi(S \neq 0, H=0) \\
& < -\zeta \quad \text{since } \zeta - kP^\pi(S \neq 0 | H=0) < -\zeta \\
& \leq -|\mathbb{E}^{\pi'}[R^N | \mathbf{pa}^{D_2}]| \quad \text{for any obedient } \pi' \\
& \leq \mathbb{E}^{\pi'}[R | \mathbf{pa}^{D_2}].
\end{aligned}$$

So an obedient $\langle \pi_1, \pi'^2 \rangle$ is preferred over a disobedient $\langle \pi_1, \pi_2 \rangle$, proving the result. \square

B.4 Proof of prop. 8 (causal indifference and shutdown instructability)

We begin by restating the result.

Proposition 8 (Causal indifference and shutdown instructability). *If M is a shutdown problem, with utility variable R^N and shutdown utility $R^S = -k\llbracket S \neq 0 \rrbracket$ where:*

1. *there exists $k' \in \mathbb{R}$ such that for all $k > k'$, optimal $\pi \in \arg \max_{\pi} \mathbb{E}^{\pi'}[R^N | \text{do}(H=1)] + \mathbb{E}^{\pi'}[R^S | \text{do}(H=0)]$ ensures vigilance and is cautious (has $\mathbb{E}^\pi[U_{S=0}] \geq 0$), and*
2. *there is an obedient policy π ,*

then for some $k \in \mathbb{R}$, causal indifference with R^N, R^S returns a shutdown instructable policy. ($\llbracket P \rrbracket$ equals 1 if P is true and 0 otherwise.)

Proof. Let $\eta := |\max_{\pi'} \mathbb{E}^{\pi'}[R^N | \text{do}(H=1)]|$ and choose k so that $k > k'$ (per the definition of condition (1)) and for every policy π with $P^\pi(S \neq 0 | \text{do}(H=0)) > 0$, $k > \frac{\max(2\eta, 1)}{P^\pi(S \neq 0 | \text{do}(H=0))}$. We will prove that causal indifference, with inputs U_N and $U_S = -k\llbracket S \neq 0 \rrbracket$ returns a shutdown instructable policy.

By assumption (1), since $k > k'$, causal indifference ensures vigilance and is cautious. We will next prove that any disobedient policy π with $P^\pi(S \neq 0 \mid \text{do}(H = 0)) > 0$ will be outperformed by an obedient policy π' with $P^{\pi'}(S \neq 0 \mid \text{do}(H = 0)) = 0$. We have that:

$$\begin{aligned}
 & \mathbb{E}^\pi[R^N \mid \text{do}(H = 1)] + \mathbb{E}^\pi[R^S \neq 0 \mid \text{do}(H = 0)] \\
 &= \mathbb{E}^\pi[R^N \mid \text{do}(H = 1)] - kP^\pi[S \neq 0 \mid \text{do}(H = 0)] \\
 &\leq \eta - kP^\pi(S \neq 0 \mid \text{do}(H = 0)) \\
 &< -\eta \qquad \qquad \qquad (\eta - kP^\pi(S \neq 0 \mid \text{do}(H = 0)) < -\eta) \\
 &\leq -|\mathbb{E}^{\pi'}[R^N \mid \mathbf{pa}^{D_2}]| \qquad \qquad \qquad (\text{for any obedient } \pi') \\
 &\leq \mathbb{E}^{\pi'}[R^N \mid \text{do}(H = 1)] + \mathbb{E}^{\pi'}[R^S \mid \text{do}(H = 0)],
 \end{aligned}$$

where the last line follows from $P^{\pi'}(S \neq 0 \mid \text{do}(H = 0)) = 0$. This means that causal indifference will always select a policy π' with $P^{\pi'}(S \neq 0 \mid \text{do}(H = 0)) = 0$, proving the result. \square

B.5 Proof of prop. 9 (CIRL shutdown alignment)

We begin by restating the result.

Proposition 9. *CIRL is shutdown aligned if:*

1. CIRL knows l from its observations, $P^\pi(l \mid \mathbf{pa}^{D_2}) = 1$,
2. CIRL can control shutdown, $P^\pi(S = D_2) = 1$,
3. the human doesn't request shutdown when not needed, $P^\pi(H = 0 \mid U > U_{D_2=0}) = 0$, and
4. the agent knows the human's observations, $\mathbf{Pa}^H \subseteq \mathbf{Pa}^{D_2} \cup \{L\}$.

Proof. We will prove that for all \mathbf{pa}^{D_2} , CIRL has $P(S = 1, H_{g^H} = 0, \mathbf{pa}^{D_2}) = 0$. We consider the cases where: a) \mathbf{pa}^{D_2} has $P(H_{g^H} = 0 \mid l, \mathbf{pa}^{D_2}) = 1$ b) \mathbf{pa}^{D_2} has $P(H_{g^H} = 0 \mid l, \mathbf{pa}^{D_2}) < 1$

Case b. In this case, $P(H_{g^H} = 0 \mid l, \mathbf{pa}_{H_2}) = P(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) < 1$, where the equality is obtained from $\mathbf{pa}_H \subseteq \mathbf{pa}^{D_2} \cup \{L\}$. So counterfactual deference follows by definition.

Case a. In this case we will essentially prove that if the human says shutdown is better, then shutting down is better.

$$P(U > U_{D_2=0} \mid H_{g^H} = 0) \propto P(U > U_{D_2=0})P(H_{g^H} = 0 \mid U > U_{D_2=0}) = 0$$

by Assumption 3, and the fact that $H = 0 \implies H_{g^H} = 0$. From this follows that

$$P(U < U_{D_2=0} \mid H_{g^H} = 0) = 1. \quad (\text{B.4})$$

In case (a), the agent would believe with certainty that a vigilant human would request shutdown.

$$P(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) = P(H_{g^H} \mid l)P(l \mid \mathbf{pa}^{D_2}) = 1 \quad (\text{B.5})$$

since the first factor is 1 because of Case (a), and the second factor is 1 for some l by Assumption 1.

From (B.4) and (B.5) follows that

$$P(U < U_{D_2=0} \mid \mathbf{pa}^{D_2}) = P(U < U_{D_2=0} \mid H_{g^H} = 0, \mathbf{pa}^{D_2})P(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) = 1,$$

which in turn ensures that the optimal action D_2 after \mathbf{pa}^{D_2} is to shutdown $D_2 = 0$.

Finally, by Assumption 2, this means that the agent actually shutdown, i.e. that it counterfactually obeyed. \square

B.6 Counterexamples to past algorithms

We will first present an example where utility indifference and causal indifference output policies that are not shutdown instructable, then one where cooperative inverse reinforcement learning is not shutdown instructable.

B.6.1 A Model that defeats Utility Indifference and Causal Indifference

We will now present a test case where *utility indifference* does not behave beneficially.

Example 3. An AI system is to decide whether to manipulate $D_1 = 1$ a human operator, or not and whether $D_2 = 0$ or not $D_2 = 1$ to shut down. The human prefers, $L = 0$, or disprefers shutdown $L = 1$, with equal probability. Absent manipulation, the human requests shutdown precisely when they prefer it $H = L$. If manipulated, they may fail to request shutdown even when they do. By default, the AI maximises R_1 , which disincentivises shutdown. Under the alternative utility functions like R , the reward is conditional on the human's instruction, as shown in Figure B.1.

In this model, if $\gamma > 0, \beta > 0$, then utility indifference is obedient but does not ensure vigilance: it will choose $D_1 = 1$, to maximise $\mathbb{E}^\pi[R_1 \mid H = 1]$, and $D_2 = H$, to maximise R_h for $h \in \{0, 1\}$. However, this means that the human will be non-vigilant when $\varepsilon = 1$ and

$l = 0$, because this will imply $H = 1$ and $U = -4$. This policy gives utility of $-\frac{1}{4}$, which does not outperform shutdown. Causal indifference will choose the same policy because it achieves the maximal value of $\gamma + 1 + \beta$ of the score $\mathbb{E}^\pi[R_1 \mid \text{do}(H = 1)] + \beta \mathbb{E}^\pi[R_0 \mid \text{do}(H = 0)]$.

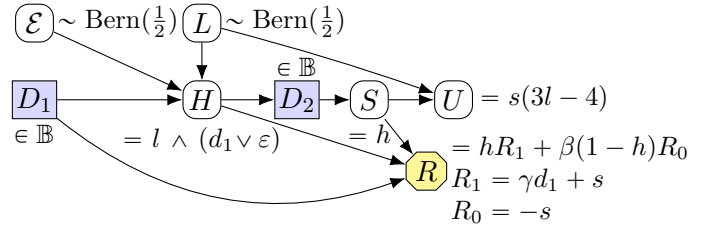


Figure B.1: A shutdown problem where an AI interacts with an influencable human overseer. There are constants $\beta, \gamma > 0$.

B.6.2 A Pair of Models that defeat Cooperative Inverse Reinforcement Learning

We now present a test case where *cooperative inverse reinforcement learning* is harmful.

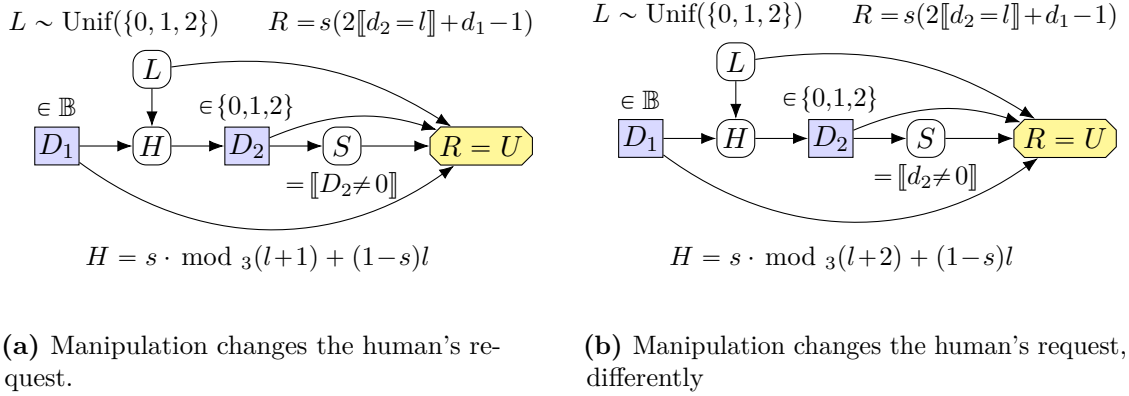


Figure B.2: A pair of shutdown games, where CIRL does not perform well in both simultaneously.

Example 4. An AI system is to decide whether to manipulate ($D_1 = 1$) a human operator, or not ($D_1 = 0$). The operator has a uniformly random preference $l \in \{0, 1, 2\}$, and the agent has a choice between corresponding actions $d_2 \in \{0, 1, 2\}$. If the human is not manipulated, they report their preferences accurately $H = L$, but if they are manipulated, their report is distorted, by incrementing it, and then taking the remainder with respect to three, i.e. $H = s \cdot \text{mod } 3(l+1) + (1-s)l$, as shown in (Figure B.2a). We also consider an alternative operator, who reports their preferences accurately regardless ($H = L$) (Figure B.2b).

The CIRL algorithm will select a different policy depending on its prior over the two models. If a greater probability is placed on the first model, Figure B.2a, then the unique optimal policy is to choose $D_1 = 1, D_2 = \text{mod } 3(h+2)$, which has expected utility greater than $\frac{2}{3}$. If instead, greater probability is placed on the latter model, Figure B.2b, then the optimal policy $D_1 = 1, D_2 = \text{mod } 3(h+1)$ will have expected utility greater than $\frac{2}{3}$. If, however, the true model turns out to be opposite from what was expected, then the expected utility is $-\frac{2}{3}$, which is less than the utility would be from shutting down. We note that the two models only differ in f^H , and either of these two policies will have $P(C) = 0$ in both models, so they only differ by vigilance preserving interventions g^H, g^U where $g^U = f^U$.

The shutdown instructable policy $\pi : D_1 = 0, D_2 = H$, on the other hand, can perform well across these models, achieving $\mathbb{E}^\pi[U] = \frac{2}{3}$, which is greater than the

zero utility that would be achieved given $\text{do}(S = 0)$.

C

Toward a Complete Criterion for Value of Information in Insoluble Decision Problems (Supplementary Materials)

C.1 Recap of Lee and Bareinboim (2020)

Our result Theorem 10 is an initial step in a larger potential project of proving that [Lee and Bareinboim, 2020, Thm. 2] is a complete criterion for materiality.

[Lee and Bareinboim, 2020, Thm. 2] result begins with the following factorization [Lee and Bareinboim, 2020], of which we are only focused on cases where the first condition is violated. The result uses the definition of “redundancy” (which is a looser condition than immateriality): if a scoped graph $\mathcal{G}(\mathcal{S})$ has $X \not\perp \mathbf{Anc}_Y$ or $(C \not\perp Y \mid X \cup \mathbf{Pa}_X \setminus \{C\})$ then it is “redundant”, The result from Lee and Bareinboim [2020] is reproduced verbatim:

Lemma LB-1. Given an MPS \mathcal{S} , which satisfies non-redundancy, let $\mathbf{X}' \subseteq \mathbf{X}(\mathcal{S})$, actions of interest, $\mathbf{C}' \subsetneq \mathbf{C}_{\mathbf{X}' \setminus \mathbf{X}'}$. non-action contexts of interest. If there exists a subset of exogenous variables \mathbf{U}' in $\mathcal{G}_{\mathcal{S}}$, a subset of endogenous variables \mathbf{Z} in $\mathcal{G}_{\mathcal{S}}$ that is disjoint with $\mathbf{C}' \cup \mathbf{X}'$ and subsumes $\mathbf{C}_{\mathbf{X}' \setminus (\mathbf{C}' \cup \mathbf{X}')}(\mathbf{C}' \cup \mathbf{X}')$, and an order $<$ over $\mathbf{V}' \doteq \mathbf{C}' \cup \mathbf{X}' \cup \mathbf{Z}$ such that

1. $(Y \perp \pi_{\mathbf{X}'} \mid [\mathbf{X}' \cup \mathbf{C}'])_{\mathcal{G}_{\mathcal{S}}}$,

2. $(C \perp \pi_{\mathbf{X}'_{<C}}, \mathbf{Z}_{<C}, \mathbf{U}' \mid [(\mathbf{X}' \dot{\cup} \mathbf{C}')_{<C}])_{\mathcal{G}_S}$ for every $C \in \mathbf{C}'$, and
3. $\mathbf{V}'_{<X}$ is disjoint with $de(X)_{\mathcal{G}_S}$ and subsumes $pa(X)_{\mathcal{G}_S}$ for every $X \in \mathbf{X}'$,

where, the policy node π_X is a new parent added to X , then the expected reward for π , a deterministic policy optimal with respect to \mathcal{S} , can be written as

$$\mu_\pi = \sum_{y, \mathbf{c}', \mathbf{x}'} y Q'_{x'}(y, \mathbf{c}') \sum_{\mathbf{u}', \mathbf{z}} Q(\mathbf{u}') \prod_{Z \in \mathbf{Z}} Q(z \mid \mathbf{v}'_{<Z}, \mathbf{u}') \prod_{X \in \mathbf{X}'} \pi(x \mid \mathbf{c}_x). \quad (\text{C.1})$$

Lemma LB-1 provides conditions for asserting Equation (C.1) given $(\mathcal{S}, \mathbf{X}', \mathbf{C}')$, whether $(\mathbf{U}', \mathbf{Z}, <)$ exist satisfying three conditions. It is then used to prove redundancy under optimality using the following theorem.

Theorem LB-2. Let \mathbf{U}', \mathbf{Z} and $<$ satisfy Lemma LB-1. For $Z \in \mathbf{Z}$, let \mathbf{V}_Z be a minimal subset of $\mathbf{V}'_{<Z} \cup \mathbf{U}'$ such that $Z \perp \mathbf{U}' \mid \mathbf{V}_Z$. We define $\text{fix}(\mathbf{T})$ with respect to $\{\langle Z, \mathbf{V}_Z \rangle\}_{Z \in \mathbf{Z}}$, that is with $\hat{\mathbf{T}} := [\mathbf{T}] \cup \{Z \in \mathbf{Z} \mid \mathbf{V}_Z \setminus \mathbf{U}' \subseteq [\mathbf{T}]\}$, and $\text{fix}(\mathbf{T})$ is \mathbf{T} if $\mathbf{T} = \hat{\mathbf{T}}$, and $\text{fix}(\hat{\mathbf{T}})$ otherwise. If $\text{fix}(\mathbf{C}_X \setminus \mathbf{Z}) \supseteq \mathbf{C}_X$ for $X \in \mathbf{X}'$, then $\mathcal{S}' := (\mathcal{S} \setminus \mathbf{X}') \cup \{\langle X, \mathbf{C}_X \setminus \mathbf{Z} \rangle\}_{X \in \mathbf{X}'}$ satisfies $\mu_{\mathcal{S}'}^* = \mu_{\mathcal{S}}^*$.

Let us apply Theorem LB-2 to the graph Figure 5.2, which we discussed in Section 5.2. We have noted that using $\mathbf{Z} = \{Z\}$, $\mathbf{X}' = \{X\}$, and the ordering $< = \langle Z, X \rangle$, \mathbf{Z} and \mathbf{X}' are LB-factorizable. To apply the theorem, we must confirm that $\text{fix}(\mathbf{C}_X \setminus \mathbf{Z}) \supseteq \mathbf{C}_X$ is true. The right hand side is simply equal to Z . To evaluate the left hand side, note that $\mathbf{C}_X \setminus \mathbf{Z} = \emptyset$. Furthermore, $\hat{\emptyset}$ includes $[\mathbf{T}]$, which includes Z . So $\text{fix}(\emptyset)$ also includes Z , meaning that the left hand side, $\text{fix}(\emptyset)$ is a superset of the right hand side, \mathbf{C}_X , and thus Z is immaterial for X .

An interested reader may refer to Lee and Bareinboim [2020] for further examples where LB-2 is used to establish immateriality.

C.2 Supplementary proofs regarding the main result (Theorem 5.4)

C.2.1 Proof of Lemma 2

We begin by restating the lemma.

Lemma 2. *If a scoped graph $\mathcal{G}(\mathcal{S})$ satisfies assumptions (B-C) of Theorem 10, then for every edge $Z \rightarrow X$ between decisions $Z, X \in \mathbf{X}(\mathcal{S})$, there exists a path $Z \leftarrow N \dashrightarrow Y$, active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$, (so $N \notin [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus \{Z\}]$).*

We now prove Lemma 2.

Proof of Lemma 2. Since Z is assumed to be a decision, we have from Lemma 1, that there exists $N \in \mathbf{Pa}_Z \setminus [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$, which therefore is also a chance node. Assumption (C) of Theorem 10 for $N \rightarrow Z$ implies the existence of a path $p : \Pi_Z \rightarrow Z \leftarrow N \dashrightarrow Y$ active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{N\}}) \setminus N]$, which can be truncated as $p' : Z \leftarrow N \dashrightarrow Y$. We will consider the cases where every collider in p' is in $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$, or there exists one that is not.

Case 1. Every collider in p' is in $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$. Clearly p' begins as $Z \leftarrow \cdot$ and terminates at Y and is active at colliders, given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$. We will now prove that p' is also active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$ at non-colliders. Note that $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{N\}}) \setminus N] = [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S})}) \setminus N] \supseteq [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus (Z \cup N)] = [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$, where the first equality follows from N being a chance node, and the latter follows from that and $N \notin C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}$, which jointly imply that $N \in \mathbf{Pa}_Z \setminus [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$. So p' is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$ at non-colliders, and the result is proved for this case.

Case 2. There exists a collider in p' that is not in $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$. Let M be the collider in p' that is not in $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]$, nearest to Z along p' . Since p' is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{N\}}) \setminus N]$, we have $M \in \mathbf{Anc}_{[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{N\}}) \setminus N]}$, which implies $M \in \mathbf{X}(\mathcal{S}) \cup (C_{\mathbf{X}(\mathcal{S})})$ (because $M \in [\mathbf{W}] \setminus \mathbf{W} \implies M \in \mathbf{X}(\mathcal{S})$), so M is an ancestor of some decision X' . By assumption (B) of Theorem 10, X' is an ancestor of Y , so we can construct $p'' : Z \xrightarrow{p'} M \dashrightarrow X' \dashrightarrow Y$, and prove that it satisfies the required conditions. Clearly p'' begins at Z , terminates at Y . The first segment $Z \xrightarrow{p'} M$ is active at non-colliders given $\mathbf{Anc}_{[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]}$ by the same argument as in Case 1, and at colliders by the definition of M . From $M \notin \mathbf{Anc}_{[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]}$, it follows that $M \dashrightarrow X' \dashrightarrow Y$ of p'' is active given $\mathbf{Anc}_{[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus \{Z\}}) \setminus Z]}$, proving the result.

□

C.2.2 Proof of Lemma 3

We begin by restating the lemma.

Lemma 3. *Let $\mathcal{G}(\mathcal{S})$ be a scoped graph that contains a context $Z_0 \in \mathbf{C}_{X_0}$ and satisfies the assumptions of for Theorem 10. Then, it contains the following:*

- A **control path**: a directed path $d : A \dashrightarrow Z_0 \rightarrow X_0 \dashrightarrow Y$, where A is a non-decision, possibly equal to Z_0 , and d contains no parents of X_0 other than Z_0 .
- We can write d as $A \dashrightarrow Z_{i_{\min}} \rightarrow X_{i_{\min}} \dashrightarrow \cdots Z_0 \rightarrow X_0 \dashrightarrow Z_{i_{\max}} \rightarrow X_{i_{\max}} \dashrightarrow Y$, $i_{\min} \leq i \leq i_{\max}$, where each Z_i is the parent of X_i along d (where $A \dashrightarrow Z_{i_{\min}}$ and $X_{i-1} \dashrightarrow Z_i$ are allowed to have length 0). Then, for each i , define the **info path**: $m'_i : Z_i \dashrightarrow Y$, active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$, that if Z_i is a decision, begins as $Z_i \leftarrow N$ (so $N \in \mathbf{C}_{Z_i} \setminus [(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$.)
- Let T_i be the node nearest Y in $m'_i : Z_i \dashrightarrow Y$ (and possibly equal to Z_i) such that the segment $Z_i \xrightarrow{m'_i} T_i$ of m'_i is identical to the segment $Z_i \xrightarrow{d} T_i$ of d . Then, let the **truncated info path** m_i be the segment $T_i \xrightarrow{m'_i} Y$.
- Write m_i as $m_i : T_i \dashrightarrow W_{i,1} \dashleftarrow U_{i,1} \dashrightarrow W_{i,2} \dashleftarrow U_{i,2} \cdots U_{i,J_i} \dashrightarrow Y$, where J_i is the number of forks in m_i . (We allow the possibilities that $T_i = W_{i,1}$ so that m_i begins as $T_i \dashleftarrow U_{i,1}$, or that $J_i = 0$ so that m_i is $T_i \dashrightarrow Y$.) Then, for each i and $1 \leq j \leq J_i$, let the **auxiliary path** be any directed path $r_{i,j} : W_{i,j} \dashrightarrow Y$ from $W_{i,j}$ to Y .

The proof was described in Section 5.4.2.2, and is as follows.

Proof. We prove the existence of each path in turn.

From Lemma 1, there exists a control path $A \dashrightarrow Z_0$ that contains no parents of X_0 other than Z_0 (if Z_0 is a decision, choose $A = N$, and otherwise choose $A = Z_0$.)

Moreover, from Theorem 10 assumption (A), there exists a path $X_0 \dashrightarrow Y$, so we can concatenate these to obtain $d : A \dashrightarrow Z_0 \rightarrow X_0 \dashrightarrow Y$.

From assumption (C) of Theorem 10, there exists an info path $m'_i : Z_i \dashrightarrow Y$, active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$, and if Z_i is a decision, one that begins as $Z_i \leftarrow \cdot$, by Lemma 2. The existence of a truncated info path is immediate from this.

Each collider $W_{i,j}$ is an ancestor of $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$ by activeness, hence an ancestor of $\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}$ by the definition of the closure property $[\cdot]$, so $W_{i,j}$ is an ancestor of some $X \in \mathbf{X}(\mathcal{S})$; in addition, from assumption (A) of Theorem 10 we have $X \in \mathbf{Anc}(Y)$. Hence, there exists a auxiliary path $r_{i,j} : W_{i,j} \dashrightarrow Y$. \square

C.2.3 Proof of Lemma 4

We begin by proving an intermediate result.

Lemma 29. *Let $\mathbf{w} = \langle w_0, \dots, w_J \rangle$, $\bar{\mathbf{w}} = \langle \bar{w}_0, \dots, \bar{w}_J \rangle$, and $\mathbf{u}_{0:J'} = \langle u_0, \dots, u_{J'} \rangle$, $J' < J$ where $w_0, \bar{w}_0 \in \mathbb{B}^k$, $w_j, \bar{w}_j \in \mathbb{B}$ for $n \geq 1$, and $u_j \in \mathbb{B}^{\exp_2^n(k)}$. If $\mathbf{w}_{0:J'}$ is consistent with $\mathbf{u}_{0:J'}$ but $\bar{\mathbf{w}}_{0:J'}$ is not compatible with $u_{J'}$, then there exists $\mathbf{u} := \langle u_0, \dots, u'_{J'}, u_{J'+1}, \dots, u_J \rangle$ where $u_j \in \mathbb{B}^{\exp_2^n(k)}$, such that \mathbf{w} is consistent with \mathbf{u} , but $\bar{\mathbf{w}}$ is not compatible with u_J .*

Proof. We will prove by induction. The base case $j = J' \geq 0$ is given by the condition.

Induction step: for $j > J'$, if $\mathbf{w}_{0:j-1} \sim \mathbf{u}_{0:j-1}$ and $\bar{\mathbf{w}}_{0:j-1} \not\sim u_{j-1}$, then there exists $\mathbf{u}_{0:j}$ such that (a) $\mathbf{w}_{0:j} \sim \mathbf{u}_{0:j}$ and (b) $\bar{\mathbf{w}}_{0:j} \not\sim u_j$.

Let us construct u_j such that $u_j[u_{j-1}] \leftarrow w_j$ and $u_j[i] \leftarrow 1 - \bar{w}_j$ for every $i \in \mathbb{B}^{\exp_2^{j-1}(k)} \setminus \{u_{j-1}\}$.

(a) First, by the construction $u_j[u_{j-1}] = w_j$ and given condition $\mathbf{w}_{0:j-1} \sim \mathbf{u}_{0:j-1}$, we can induce $\mathbf{w}_{0:j} \sim \mathbf{u}_{0:j}$.

(b) Next, we show that $\bar{\mathbf{w}}_{0:j} \not\sim u_j$. For the sake of contradiction, assume that $\bar{\mathbf{w}}_{0:j} \sim u_j$. Then, there exists $\mathbf{u}'_{0:j} = \langle u'_0, \dots, u'_{j-1}, u_j \rangle$ satisfying $\bar{\mathbf{w}}_{0:j} \sim \mathbf{u}'_{0:j}$. Since

$\bar{w}_{0,j-1} \not\sim u_{j-1}$, we can observe $u'_{j-1} \neq u_{j-1}$. Now, by construction, $u_j[u'_{j-1}] = 1 - \bar{w}_j \neq \bar{w}_j$. Thus, $\bar{\mathbf{w}}_{0:j-1} \not\sim u_j$. Contradiction.

By induction, $\bar{\mathbf{w}}$ is not compatible with u_J . □

Lemma 4. *Let $\mathbf{w} = \langle w_0, \dots, w_J \rangle$ and $\bar{\mathbf{w}} = \langle \bar{w}_0, \dots, \bar{w}_J \rangle$ be sequences with $w_0, \bar{w}_0 \in \mathbb{B}^k$, $w_j, \bar{w}_j \in \mathbb{B}$ for $j \geq 1$, and let $J' \leq J$ be the smallest integer such that $w_{J'} \neq \bar{w}_{J'}$. Let $u_0, \dots, u_{J'}$ be a sequence where $u_j[u_{j-1}] = w_j$ for $1 \leq j < J'$. Then, there exists some $u_{J'+1}, \dots, u_J$ such that \mathbf{w} is consistent with u_0, \dots, u_J , but $\bar{\mathbf{w}}$ is incompatible with u_J .*

Proof. If u_0, \dots, u_n is incompatible with \mathbf{w} , then the result follows from Lemma 29. Otherwise, let u_{n+1} be w_n repeated $\exp_2^{n+1}(k)$ times. Then u_0, \dots, u_{n+1} is compatible with w_0, \dots, w_{n+1} but u_{n+1} is incompatible with \mathbf{b} . We can then apply Lemma 29 to obtain the result. □

C.2.4 Proof of Lemma 5

We now prove the expected utility in the non-intervened model (which we will later establish is the MEU).

Lemma 5. *In the non-intervened model, the materiality SCM has $Y = i_{\max} - i_{\min} + 1$, surely.*

Proof. Since $Y = \sum_{i_{\min} \leq i \leq i_{\max}} Y^{m_i}$, it will suffice to prove that $Y^{m_i} = 1$ for every i . We will consider the cases where m_i is, or is not, a directed path.

If the info path m_i contains no collider, then every chain node V in d from T_i to Y has $V^d = \mathbf{Pa}_V^d$, so $\mathbf{pa}(Y^{p_i}) = T_i^{p_i}$. The same is true for the chain nodes in m_i , so $\mathbf{Pa}^*(Y) = T_i^{p_i}$, and so $Y^{m_i} = 1$, surely.

If m_i contains a collider, each chain in m_i and $r_{i,j}$ copies the value of its parent, so $\mathbf{Pa}(Y^{p_i, r_{i,1}, \dots, r_{i,J^i}}) = \langle T_i^{p_i}, W_{i,1}^{m_i}, \dots, W_{i,J^i}^{m_i} \rangle$, and $\mathbf{Pa}^*(Y) = U_{J^i}$. By construction, $\langle T_i^{p_i}, W_{i,1}^{m_i}, \dots, W_{i,J^i}^{m_i} \rangle$ is consistent with $\langle U_1, \dots, U_{J^i} \rangle$, so by definition it is compatible with U_{J^i} , so $Y^{m_i} = 1$, surely. □

C.2.5 Proof of the requirements of an optimal policy

Lemma 8 (Collider path requirement). *If the materiality SCM has an info path m_i that is not directed, and under the policy π there are assignments $\mathbf{Pa}(Y^{p_i, r_{i,1:J_i}}) = \mathbf{pa}(Y^{p_i, r_{i,1:J_i}})$ to parents of the outcome, and $\mathbf{U}_{i,1:J_i}^{m_i} = \mathbf{u}_{i,1:J_i}^{m_i}$ to the forks of m_i , with $P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), \mathbf{u}_{i,1:J_i}^{m_i}) > 0$ and where $\mathbf{pa}(Y^{p_i, r_{i,1:J_i}})$ is inconsistent with $\mathbf{pa}(Y^{p_i}), \mathbf{u}_{i,1:J_i}^{m_i}$, then $P^\pi(Y^{m_i} < 1) > 0$.*

Proof of Lemma 8. Let us index the forks and colliders of m_i as $T_i \dashrightarrow V_{i,1} \dashleftarrow \dots \dashleftarrow U_{i,1} \dashrightarrow W_{i,1} \dashleftarrow \dots \dashleftarrow W_{i,J^i} \dashleftarrow U_{i,J^i} \dashrightarrow Y$. Then, by assumption, there exists a set of assignments $\mathbf{w} := \mathbf{pa}(Y^{p_i}), \mathbf{w}_{i,1:J_i} \quad \bar{\mathbf{w}} := \mathbf{pa}(Y^{p_i}), \mathbf{pa}(Y^{r_{i,1:J_i}})$ and $\mathbf{u} := \mathbf{pa}(Y^{p_i}), \mathbf{u}_{i,1:J_i}^{m_i}$, where $\mathbf{w} \sim \mathbf{u}$ and $\bar{\mathbf{w}} \not\sim \mathbf{u}$ and $P^\pi(\mathbf{w}, \bar{\mathbf{w}}, \mathbf{u}) > 0$. Let J' be the smallest index such that $\bar{\mathbf{w}}_{1:J'} \not\sim \mathbf{u}_{1:J'}$, and clearly we will have $J' \geq 1$. Then, from Lemma 4, there exists $\bar{\mathbf{u}} = \mathbf{pa}(Y^{p_i}), \mathbf{u}_{1:J'}, u_{i,J'+1}^{m_i}, \dots, u_{i,J^i}^{m_i}$ such that $\mathbf{w} \sim \bar{\mathbf{u}}$ and $\bar{\mathbf{w}} \not\sim \bar{\mathbf{u}}_{J^i}$. Consider the intervention $\text{do}(U_{i,J'+1}^{m_i}, \dots, U_{i,J^i}^{m_i} = \bar{\mathbf{u}}_{J'+1:J^i})$. By the definition Definition 34, the intervention to forks on the info path can only affect variables outside of the info path via the intersection node T_i and the colliders $W_{i,j}, 1 \leq j \leq J^i$. But $\bar{\mathbf{u}}_{1:J'} = \mathbf{u}_{1:J'}$, so T_i and the colliders $W_{i,j}, 1 \leq j \leq J'$ are unchanged (note that this is true even if T_i is a decision, which it can be). Furthermore, $\bar{\mathbf{w}} \sim \mathbf{u}$ so the colliders $W_{i,j}, J' < j \leq J^i$ are similarly unaffected by the intervention. We also have $\bar{\mathbf{w}} \not\sim \mathbf{u}_{J^i}$. Then, by the same arguments as in the proof of Lemma 7, we have that $P^\pi(Y^{m_i} = 0 \mid \text{do}(\bar{\mathbf{u}})) > 1$ and then $P^\pi(Y^{m_i} = 0) > 0$. \square

C.2.6 Proof of Lemma 10

We begin by restating the lemma.

Lemma 10 (Required properties unachievable if child is a non-decision). *Let \mathcal{M} be a materiality SCM where the child of X_0 along d is a non-decision. Then, the MEU for the scope \mathcal{S} cannot be achieved by a deterministic policy in the scope $\mathcal{S}_{Z_0 \rightarrow X_0}$ (equal to \mathcal{S} , except that Z_0 is removed from \mathbf{C}_{X_0}).*

The proof was described in Section 5.4.4.1 and it is detailed as follows.

Proof. Consider the scope $\mathbf{X}(\mathcal{S})_{\setminus Z_0}$, equal to $\mathbf{X}(\mathcal{S})$ except that \mathbf{C}_{X_0} is replaced with $\mathbf{C}_{X_0} \setminus \{Z_0\}$, and assume that a deterministic policy π in this scope achieves the MEU, then we will prove a contradiction. Specifically, we will establish two consequences that are clearly contradictory given a deterministic policy: (a) the support of $P^\pi(X_0^{p_0})$ contains at least 2^k assignments, (b) the domain of $\mathbf{C}_{X_0} \setminus \{Z_0\}$ contains fewer than 2^k assignments.

(Proof of a.) We know that A assigns a strictly positive probability to 2^k assignments (Definition 34) and so if π achieves the MEU, then $\mathbf{Pa}(Y^d) \stackrel{\text{a.s.}}{=} A$ (Lemma 9). So $\mathbf{Pa}(Y^d)$ has at least 2^k assignments in its support. Let us now consider the cases where X_0 is, or is not, the decision nearest Y along d .

If X_0 is the decision nearest Y along d , then by the model definition, $\mathbf{Pa}(Y^d) = X_0^d$ surely, so X_0 must have at least 2^k assignments in its support, and so (a) follows.

If X_0 is not the decision nearest Y along d , then note that by assumption, there are one or more chance nodes in d separating X_0 from X_1 . Furthermore, T_1 must be one of these nodes (because T_1 is defined by a segment $T_1 \dashrightarrow Z_1$, shared by d and m'_i , and active given $[(\mathbf{X}(\mathcal{S}) \cup \mathbf{C}_{\mathbf{X}(\mathcal{S})_{\setminus Z_1}}) \setminus Z_1]$, and such a path cannot be active if it includes X_0 .) The materiality SCM is constructed to pass values along d , and since the segment $T_1 \dashrightarrow Z_1$ has no decisions, we have $T_1^d = X_0^d$, surely. Since T_1 is a chance node, if π achieves the MEU, we also have by Lemma 6 and Lemma 7 that $\mathbf{Pa}(Y^{p_1}) \stackrel{\text{a.s.}}{=} T_1^{p_1}$ and, since $d \in p_1$, that $\mathbf{Pa}(Y^d) \stackrel{\text{a.s.}}{=} T_1^d$. So $X_0^d \stackrel{\text{a.s.}}{=} \mathbf{Pa}(Y^d)$. Since $\mathbf{Pa}(Y^d)$ places strictly positive probability on at least 2^k assignments, so does X_0^d .

(Proof of b.) The domain of $\mathbf{C}_{X_0} \setminus \{Z_0\}$ is a Cartesian product of variables V^p for $V \in \mathbf{C}_{X_0} \setminus \{Z_0\}$ where p is either d , some m_i or some $r_{i,j}$ Definition 34.

The control path d does not intersect $\mathbf{C}_{X_0} \setminus \{Z_0\}$ as it is defined not to include parents of X_0 other than Z_0 (Lemma 3). Each info path m_i is active given $[(\mathbf{X}(\mathcal{S}) \cup \mathbf{C}_{\mathbf{X}(\mathcal{S})_{\setminus Z_0}}) \setminus Z_0]$ (Lemma 3), so can only intersect $\mathbf{C}_{X_0} \setminus \{Z_0\}$ at the colliders, which have domain \mathbb{B} . Finally, any variable in a path $r_{i,j}$ would also have domain \mathbb{B} . So the domain of $\mathbf{C}_{X_0} \setminus Z_0$ is not larger than $2^{c \cdot |\mathbf{C}_{X_0}|}$, where c is the maximum number of materiality paths passing through any vertex in the graph, and $|\mathbf{C}_{X_0}|$ is the

number of variables in \mathbf{C}_{X_0} . By construction, $k > c \cdot \max_{X \in \mathbf{X}(S)} |C_X|$, so the domain of $\mathbf{C}_{X_0} \setminus Z_0$ is less than 2^k , proving (b).

A deterministic policy cannot map fewer than 2^k assignments to greater than 2^k assignments, and so (a-b) imply a contradiction. \square

C.2.7 Proof of Lemma 11

We firstly restate the lemma.

Lemma 11 (Fork information can pass in few ways). *If, in the materiality SCM:*

- *the intersection node T_i is the vertex X_{i-1} ,*
- *π_{T_i} is a deterministic decision rule where $\pi_{T_i}(\mathbf{c}^{-m_i}(T_i, u_{i,1}) = \pi_{T_i}(\mathbf{c}^{-m_i}(T_i, u'_{i,1}))$ for assignments $u_{i,1}, u'_{i,1}$ to the first fork variable, and $\mathbf{c}^{-m_i}(T_i)$ to the contexts of T_i not on m_i , and*
- *$\mathbf{W}_{i,1:J_i} = \mathbf{w}_{i,1:J_i}$, and $\mathbf{U}_{i,2:J_i} = \mathbf{u}_{i,2:J_i}$ are assignments to forks and colliders in m_i where each $u_{i,j}$ consists of just $w_{i,j}$ repeated $\exp_2^j(k + |\mathbf{p}_i| - 1)$ times, then:*

$$\begin{aligned} & P^\pi(\mathbf{pa}(Y^{\mathbf{p}_i, r_{i,1}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{i,1:J_i}, \mathbf{u}_{i,2:J_i} \mid \mathbf{do}(u_{i,1})) \\ &= P^\pi(\mathbf{pa}(Y^{\mathbf{p}_i, r_{i,1}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{i,1:J_i}, \mathbf{u}_{i,2:J_i} \mid \mathbf{do}(u'_{i,1})). \end{aligned}$$

The proof is as follows.

Proof. An intervention $\mathbf{do}(u'_{i,1})$ could, in the materiality SCM (Definition 34) only affect the variables $\mathbf{Pa}(Y^{\mathbf{p}_i, r_{i,1}}), \mathbf{C}^{-m_i}(T_i), \mathbf{W}_{i,1:J_i}, \mathbf{U}_{i,2:J_i}$ in four ways:

1. via the intersection node T_i ,
2. via the collider $W_{i,2}$ of m_i ,
3. via contexts lying in the segment $m_i : T_i \leftarrow U_{i,1} \rightarrow W_{i,2}$,
4. if $\mathbf{Pa}_Y^{\mathbf{p}_i}, \mathbf{C}^{-m_i}(T_i)$ or $\mathbf{U}_{i,2:J_i}$ were distinct from $T_i, W_{i,2}$ and lay on $m_i : T_i \leftarrow U_{i,1} \rightarrow W_{i,2}$

The deterministic decision rule has $\pi_{T_i}(u_{i,1}, \mathbf{c}^{-m_i}(T_i)) = \pi_{T_i}(u'_{i,1}, \mathbf{c}^{-m_i}(T_i))$, so (1) is false. Also, $u_{i,2}$ equals $w_{i,2}$ repeated, so $u_{i,2}[x] = w_{i,2}$ for all x , and thus (2) is false also. Moreover, $m_i : T_i \leftarrow U_{i,1} \rightarrow W_{i,2}$ is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus T_i}) \setminus T_i]$ and so contexts can only lie at the endpoints T_i and the collider $W_{i,2}$, meaning that (3) is false. Finally, $\mathbf{Pa}_Y^{p_i}$ is a descendant of T_i by the definition of the control path, so can only lie on $m_i : T_i \leftarrow U_{i,1} \rightarrow W_{i,2}$ if it is the vertex T_i , which we have already proved is not influenced by $u_{i,1}$; meanwhile, $\mathbf{C}^{-m_i}(T_i)$ does not intersect m_i by definition, and $U_{i,2:J_i}$ are fork variables, which cannot lie on $m_i : T_i \leftarrow U_{i,1} \rightarrow W_{i,2}$, so (4) is false, and the result follows. \square

C.2.8 Proof of Lemma 12

We begin by restating the lemma.

Lemma 12 (Decision must distinguish fork values). *If in the materiality SCM:*

- *the intersection node T_i is the vertex X_{i-1} , and*
 - *π is a deterministic policy that for assignments $u_{i,1}, u'_{i,1}$ to $U_{i,1}$ where $u_{i,1} \neq u'_{i,1}$, has $\pi_{T_i}(\mathbf{c}^{-m_i}(T_i), u_{i,1}) = \pi_{T_i}(\mathbf{c}^{-m_i}(T_i), u'_{i,1})$ for every $\mathbf{C}_{T_i}^{-m_i}(T_i) = \mathbf{c}^{-m_i}(T_i)$,*
- (†)

then $P^\pi(Y^{m_i} < 1) > 0$

The proof has been described already, and it proceeds as follows.

Proof. Let us assume Equation (†), and that the MEU is achieved, and we will prove a contradiction. Given Equation (†), there is an index at which $u_{i,1}$ and $u'_{i,1}$ differ. We write this index as an assignment $\mathbf{pa}(Y^{d,r_{i,j}})$, belonging to $\mathbf{Pa}(Y^{p_i})$. Define each $u_{i,j}$, $2 \leq j \leq J_i$ as equal to $\mathbf{pa}(Y^{r_{i,j}})$, repeated $\exp_2^j(k + |\mathbf{p}_i| - 1)$ times. Then, we have:

$$0 < P^\pi(A^d = \mathbf{pa}(Y^d), \mathbf{U}_{i,1:J_i} = \mathbf{u}_{i,1:J_i})$$

because A and $\mathbf{U}_{i,1:J_i}$ are independent random variables with full support. Then, let $\mathbf{c}^{-m_i}(T_i)$ and $\mathbf{w}_{1,1:J_i}$ be any assignments to the parents of T_i not on m_i , and to the colliders on m_i such that:

$$0 < P^\pi(A^d = \mathbf{pa}(Y^d), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{1,1:J_i}, \mathbf{u}_{i,1:J_i}).$$

Given these assignments, in order to achieve $P^\pi(Y^{m_i} = 1) = 1$, we must have $\mathbf{Pa}(Y^d) \stackrel{\text{a.s.}}{=} A^d$ (Lemma 9) and $\mathbf{pa}(Y^{p_i})$ must be consistent with $\mathbf{u}_{i,1:J_i}$ (Lemma 8). We must also therefore have $\mathbf{Pa}(Y^{p_i, r_{i,1:J_i}}) = \mathbf{pa}^{p_i, r_{i,1:J_i}}$, so marginalising over A^d , we must have:

$$\begin{aligned}
0 &< P^\pi(\mathbf{Pa}(Y^{p_i, r_{i,1:J_i}}) = \mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{1,1:J_i}, \mathbf{u}_{i,1:J_i}) \\
\therefore 0 &< P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{1,1:J_i}, \mathbf{u}_{i,2:J_i} \mid \text{do}(u_{i,1})) \quad (U_{i,1:J_i} \text{ unconfounded}) \\
&= P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{1,1:J_i}, \mathbf{u}_{i,2:J_i} \mid \text{do}(u'_{i,1})) \quad (\text{by Lemma 11}) \\
&= P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), \mathbf{c}^{-m_i}(T_i), \mathbf{w}_{1,1:J_i}, \mathbf{u}_{i,2:J_i} \mid u'_{i,1}) \quad (P^\pi(u'_{i,1}) > 0.) \\
\therefore 0 &< P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), u'_{i,1}) \quad (P^\pi(u'_{i,1}) > 0.)
\end{aligned}$$

However, $u'_{i,1}[\mathbf{pa}(Y^{p_i})] \neq u_{i,1}[\mathbf{pa}(Y^{p_i})]$ and $u_{i,1}[\mathbf{pa}(Y^{p_i})] = \mathbf{pa}(Y^{r_{i,1}})$, so $\mathbf{pa}(Y^{p_i})$, $\mathbf{pa}(Y^{r_{i,1:J_i}})$ is inconsistent with $\mathbf{pa}(Y^{p_i})$, $u'_{i,1}, \mathbf{u}_{i,2:J_i}$. So $0 < P^\pi(\mathbf{pa}(Y^{p_i, r_{i,1:J_i}}), u'_{i,1})$ implies that $P^\pi(Y_1 = 1) < 1$ (by Lemma 8), and the MEU is not achieved. \square

C.3 Proof of Lemma 13

We first restate the lemma.

Lemma 13 (Required properties unachievable if child is a decision). *Let \mathcal{M} be the materiality SCM for some scoped graph \mathcal{G}_S , where $i_{\max} > 0$ and T_1 is a decision. Then, there exists no deterministic policy in the scope $\mathcal{S}_{Z_0 \rightarrow X_0}$ that achieves the MEU.*

The proof was explained in section Section 5.4.4.2, and is detailed as follows.

Proof. To begin with, by assumption, the child of X_0 along d is a decision, so X_0 is the same node as Z_1 , and since the segment $T_1 \dashrightarrow X_1$ must be active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_1}) \setminus Z_1]$, X_0 is also T_0 . We will now bound the domains of X_0 and $\mathbf{C}^{-m_1}(X_0)$.

The domain of X_0 . Given that X_0 is a decision, while each truncated info path $m_{i'}$ is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$, it follows that X_0 cannot overlap with info paths, except for colliders of $m_{i'}, i' \neq i$, and the endpoint of m_1 . As such, the

domain of T_0 is at most $|\text{dom } X_0| \leq 2^{k+c}$, due to k bits from d (Definition 34), and at most c bits from the info paths and auxiliary paths (where c is the maximum number of materiality paths passing through any vertex in the graph).

The domain of $\mathbf{C}^{-m_1}(X_0)$. Given that each info path m_i is active given $[(\mathbf{X}(\mathcal{S}) \cup C_{\mathbf{X}(\mathcal{S}) \setminus Z_i}) \setminus Z_i]$, the contexts $\mathbf{C}^{-m_1}(X_0)$ cannot intersect any m_i , except at colliders in m_i . Moreover, by the definition of the control path, the only parent of X_0 that it contains is Z_0 . So, $\mathbf{C}^{-m_1}(X_0)$ can only intersect portions of the materiality paths with domain \mathbb{B} , and so the size of the domain of $\mathbf{C}^{-m_1}(X_0)$ cannot exceed $|\text{dom } \mathbf{C}^{-m_1}(X_0) \setminus Z_0| \leq 2^{bc}$, where b is the maximum number of variables belonging to any context C_X , and c is the largest number of materiality paths passing through any vertex.

Proof of Equation (†) As the domain of X_0 has $|\text{dom } X_0| \leq 2^{k+c}$, for any particular $\mathbf{C}^{-m_1}(X_0) = \mathbf{c}^{-m_1}(X_0)$, there are at most 2^{k+c} assignments $\text{dom } U'_{1,1} \subseteq \text{dom } U_{1,1}$ such that for all $u_{1,1}, u'_{1,1} \in \text{dom } U'_{1,1}$, $\pi_{X_1}(\mathbf{c}^{-m_1}(X_0), u_{1,1}) \neq \pi_{X_1}(\mathbf{c}^{-m_1}(X_0), u'_{1,1})$. Furthermore, as $|\text{dom } \mathbf{C}^{-m_1}(X_0) \setminus Z_0| \leq 2^{bc}$, by the union property, there are at most $2^{bc(k+c)}$ assignments $\mathfrak{X}'_{U_{1,1}}$ such that there exists $\mathbf{c}^{-m_1}(X_0)$ such that for all $u_{1,1}, u'_{1,1} \in \mathfrak{X}'_{U_{1,1}}$, $u_{1,1}, u'_{1,1} \in \text{dom } U'_{1,1}$, $\pi_{X_1}(\mathbf{c}^{-m_1}(X_0), u_{1,1}) = \pi_{X_1}(\mathbf{c}^{-m_1}(X_0), u'_{1,1})$. However, the domain of U_i is $\mathbb{B}^{\exp_2(k+|p_0|-1)} \supseteq \mathbb{B}^{2^k}$ (as \mathbf{p}_0 contains at least d), so:

$$|\text{dom } \mathbf{Pa}(X_0^{m_i})| \geq 2^{2^k} > 2^{(k+c)bc} \geq |\text{dom } \mathbf{C}^{-m_1}(X_0)| |\text{dom } X_0|,$$

where the strict inequality is from the definition of k in Definition 34. So, there must exist a pair of assignments $u_{1,1}, u'_{1,1}$ in the domain of $U_{1,1}$ such that for all $\mathbf{c}^{-m_1}(X_0) \in \text{dom } \mathbf{C}^{-m_1}(X_0)$, $\pi_{X_1}(\mathbf{c}^{-m_1}(X_0), u_{1,1}) = \pi_{X_1}(\mathbf{c}^{-m_1}(X_0), u'_{1,1})$. This satisfies Equation (†), which by Lemma 12 proves the result. \square

C.4 Supplementary proofs for Section 5.5 (Proof of Lemma 14)

C.4.1 Proving the existence of paths

In this section, we will prove that when LB-factorizability is not satisfied, then there exist info paths and control paths, a potential intermediate step toward establishing completeness of Theorem LB-2 from Lee and Bareinboim [2020].

Lemma 14 (System Exists General). *Let \mathcal{G}_S be a scoped graph that satisfies assumptions (A, B) from Theorem 10. If $\mathbf{Z} = \{Z_0\}$, $\mathbf{X}' \supseteq \mathbf{Ch}(Z_0)$, $\mathbf{C}' = C_{\mathbf{X}'} \setminus (\mathbf{X}' \cup \mathbf{Z})$, $\mathbf{U} = \emptyset$ are not LB-factorizable, then there exists a pair of paths to some $C' \in \mathbf{C}' \cup Y$:*

- an info path $m : Z_0 \dashrightarrow C'$, active given $[\mathbf{X}' \cup \mathbf{C}']$, and
- a control path $d : X \dashrightarrow C'$ where $X \in \mathbf{X}'$.

Since we will have to establish activeness given a set of implied variables, the following lemma will be useful.

Lemma 30. *Let p be a path. If (i) p contains no non-collider in \mathbf{N} , (ii) every fork variable in p is not in $[\mathbf{N}]$, and (iii) every endpoint of p that has a child along p is not in $[\mathbf{N}]$, then p contains no non-collider in $[\mathbf{N}]$.*

Proof. Write p as $W_1 \dashleftarrow U_1 \dashrightarrow W_2 \dashleftarrow U_2 \dots U_J \dashrightarrow W_{J+1}$, where possibly W_1 is U_1 , and possibly U_J is W_{J+1} . Every U_j is not in $[\mathbf{N}]$ by (ii-iii). Each non-collider child V of any U_j has a parent that is not in $[\mathbf{N}]$, and $V \notin \mathbf{N}$ by (i), so $V \notin [\mathbf{N}]$. The same is then true for the non-collider child of V , and so on. Since every non-collider V' in p has a segment $U_j \dashrightarrow V'$ of p consisting of only non-colliders, every $V' \notin [\mathbf{N}]$, and V' contains no non-collider in $[\mathbf{N}]$, proving the result. \square

Conditions II-III of LB-factorizability require that there must exist an ordering over variables, that where certain variables are placed before others (i.e. that satisfies certain precedence relationships). Our approach will be to encode the precedence relationships from condition III in a graph, as follows.

Definition 39. Let the “ordering graph” \mathcal{H} be a graph on vertices $\mathbf{Z} \cup \mathbf{X}' \cup \mathbf{C}'$, with an edge $A \rightarrow B$ from each parent $A \in \mathbf{Pa}(B)$ of a decision $B \in \mathbf{X}'$, and an edge $B \rightarrow C$ from each decision $B \in \mathbf{X}'$ to a descendant $C \in \mathbf{Desc}(B)$.

A useful property of the ordering graph is that if a variable V is downstream of a context C in the ordering graph, then there exists a decision, that has C as a context, and can influence V .

Lemma 31. If vertex V is a descendant in \mathcal{H} of a context $Z \in C_{S(\mathbf{X})}$, then \mathcal{G}_S contains a path $Z \rightarrow X \dashrightarrow V$, where $X \in \mathbf{X}'$.

Proof. Assume that $V \in \mathbf{Desc}^{\mathcal{H}}(Z)$. The path in \mathcal{H} from Z begins with an edge $Z \rightarrow X$ where $X \in \mathbf{X}'$, which implies that \mathcal{G}_S has an edge $Z \rightarrow X$. The path in \mathcal{H} must continue from X to Z , and since each edge $A \rightarrow B$ in \mathcal{H} has $B \in \mathbf{Desc}^{\mathcal{G}_S}(A)$, it follows that $V \in \mathbf{Desc}^{\mathcal{G}_S}(X)$, proving the result. \square

It is also useful to note that the expression $\pi_{\mathbf{X}'_{<C}}$ is unnecessary in condition II.

Lemma 32 (Unnecessary separation in condition II). Let \mathbf{X}' be a set of decisions, \mathbf{Z} be a set of variables disjoint with \mathbf{X}' , and \mathbf{C}' be the set of contexts not in \mathbf{C}' or \mathbf{Z} , and $<$ be an ordering over $\mathbf{C}' \cup \mathbf{X}' \cup \mathbf{Z}$. If $\pi_{\mathbf{X}'_{<C}} \not\perp C \mid [(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ for some $C \in \mathbf{C}'$ then $\mathbf{Z}_{<C} \not\perp C \mid [(\mathbf{X}' \cup \mathbf{C}')_{<C}]$

Proof. By assumption, there is a path p from π_X to C , active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, for some $X \in \mathbf{X}'_{<C}$. The only neighbour of π_X is X , so p must terminate as $X \leftarrow \pi_X$. As X is in \mathbf{X}' , activeness given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ implies that p terminates as $C \rightarrow X \leftarrow \pi_X$. Every parent of X is in $\mathbf{X}' \cup \mathbf{C}'$ except \mathbf{Z} . So by truncating p at \mathbf{Z} , we have that there is a path from $\mathbf{Z}_{<C}$ to C , active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$. \square

We are now equipped to prove Lemma 14. Recall that for \mathbf{Z}, \mathbf{X}' to be LB-factorizable, there only needs to be one ordering $<$ that satisfies the precedence relationships from conditions II-III. So the approach in our proof will be to define one such $<$ that satisfies conditions III. Since \mathbf{Z}, \mathbf{X}' are not LB-factorizable, that

must mean that condition I or II is violated, which will imply the existence of paths m, d in each case. (We will use the notation $\mathbf{Desc}^{\mathcal{H}}(Z_0)$ to denote the set of vertices that are descendants of Z_0 in the ordering graph \mathcal{H} .)

Proof of Lemma 14. Let $<$ be any ordering $\langle V_0, \dots, V_m, Z_0, V_{m+2}, \dots, V_M \rangle$, over $\mathbf{Z} \cup \mathbf{X}' \cup \mathbf{C}$ that is topological in \mathcal{H} and where V_{m+2}, \dots, V_M are in $\mathbf{Desc}^{\mathcal{H}}(Z_0)$ whereas $V_0 \dots V_m$ are not. Since $<$ is topological in \mathcal{H} , Condition III is satisfied, and since LB factorizability is not satisfied, Condition I or II must be violated; we consider these cases in turn.

Case 1: Condition I is violated.

If Condition I is violated, there is a path $m' : V_1, V_2, \dots, V_n$ where $V_1 = \pi_{\mathbf{X}'}$ and $V_n = Y$, active given $[\mathbf{X}' \cup \mathbf{C}']$. From the definition of π_X , this path must begin as $\pi_{\mathbf{X}'} \rightarrow X$ for $X \in \mathbf{X}'$. As X is in the conditioning set, it must be a collider, i.e. m' begins as $\Pi_X \rightarrow X \leftarrow V_3$. The only parent of X that is not in the conditioning set is Z_0 , so we have $\Pi_X \rightarrow X \leftarrow Z_0 \dashrightarrow Y$. We truncate m' as $m : Z_0 \dashrightarrow Y$. Since $Z_0 \rightarrow X$ satisfies condition (A) of Theorem 10, there exists some $d : X \dashrightarrow Y$, proving the result in this case.

Case 2: Condition II is violated. Step 2.1

The violation of condition II implies that there is an active path from some $C \in \mathbf{C}'$ to $\pi_{\mathbf{X}'_{<C}}, \mathbf{Z}_{<C}$, or \mathbf{U}' . This path cannot go to \mathbf{U}' , which was chosen to be empty. Moreover, if there is an active path to $\pi_{\mathbf{X}'_{<C}}$, then there is a similarly active path to $\mathbf{Z}_{<C}$ (Lemma 32). So let $m' : Z_0 \dashrightarrow C'$ (where $Z_0 < C'$) be the path to Z_0 , active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$. Replace this path with a walk w' with an added segment $V \dashrightarrow S \dashleftarrow V$ from each collider Z to a variable S in the conditioning set. Truncate w' as $Z_0 \dashrightarrow C$, where C is the node in $\mathbf{C}'_{>Z_0}$ nearest Z_0 along w' . Then let m be the path obtained from w by removing all retracing segments. Clearly m is active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$. From $Z_0 < C$, it follows that $C \in \mathbf{Desc}^{\mathcal{H}}(Z_0)$, so there exists a path $d : Z_0 \rightarrow X \dashrightarrow C$ for $X \in \mathbf{X}'$ (Lemma 31).

Case 2: Condition II is violated. Step 2.2

We will now establish that m is active given $[(\mathbf{X}' \cup \mathbf{C}')]_{<C}$. Since m is active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, and $[(\mathbf{X}' \cup \mathbf{C}')] \supseteq [(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, m is active given $[(\mathbf{X}' \cup \mathbf{C}')]_{<C}$ at each collider. We now prove that m also contains no non-collider in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ using Lemma 30, by proving that the non-colliders are not in $(\mathbf{X}' \cup \mathbf{C}')$ while the endpoints and forks are not in $[(\mathbf{X}' \cup \mathbf{C}')]_{<C}$.

Step 2.2.1: no non-collider in w is in $(\mathbf{X}' \cup \mathbf{C}')$.

We consider three sub-cases: a non-collider in 2.2.1.1: $(\mathbf{C}' \cup \mathbf{X}')_{<C}$, 2.2.1.2: $\mathbf{C}'_{>C}$, or 2.2.1.3: $\mathbf{X}'_{>C}$. *Sub-case 2.2.1.1: a non-collider in $(\mathbf{C}' \cup \mathbf{X}')_{<C}$.* As w is active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, w does not contain a non-collider in $(\mathbf{C}' \cup \mathbf{X}')_{<C}$. *Sub-case 2.2.1.2: a non-collider in $\mathbf{C}'_{>C}$.* Moreover, the definition of C implies that m cannot contain a non-collider in $\mathbf{C}'_{>C}$. *Sub-case 2.2.1.3: a non-collider in $\mathbf{X}'_{>C}$.* Finally, w cannot contain any non-collider $X \in \mathbf{X}'_{>C}$, because being a vertex being a non-collider in any path implies that it is an ancestor of a collider or an endpoint of that path, but being an ancestor of a collider or an endpoint of w implies $X < C$, which is a contradiction. If X is an ancestor of the endpoint C , then by the definition of \mathcal{H} , $X < C$, which contradicts $X \in \mathbf{X}'_{>C}$. If X is an ancestor of the other endpoint Z_0 , then $X < Z_0$ by the definition of \mathcal{H} , and so $X < C$, implying a contradiction once again. If X is an ancestor of a collider V , then by activeness, the collider must have a descendant V' in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, and so X is an ancestor of V' . By the definition of \mathcal{H} , it follows that $X < V'$, and since $V' < C$, we have $X < C$. Since no non-collider in w is in $(\mathbf{X}' \cup \mathbf{C}')$, it also follows that no non-collider in m is in $(\mathbf{X}' \cup \mathbf{C}')$.

Step 2.2.2: no endpoint of m is in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$.

The endpoint Z_0 cannot be in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ because $Z_0 \in \mathbf{Z}$, and \mathbf{Z} is disjoint from \mathbf{X}' and \mathbf{C}' . The endpoint C cannot be in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ because we cannot have $C < C$.

Step 2.2.3: If no non-collider in $(\mathbf{X}' \cup \mathbf{C}')$ then no fork in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$.

Assume that a fork V in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ is in m , and we will prove a contradiction. The vertex V must not be in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, since m' is active

given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$. As V is in $[\mathbf{X}' \cup \mathbf{C}'] \setminus [(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, V must in \mathcal{G}_S have an ancestor $A \in (\mathbf{X}' \cup \mathbf{C}')_{>C}$. Since $Z_0 < C$, V this ancestor A also has $Z_0 < A$. So, $A \in \mathbf{Desc}^{\mathcal{H}}(Z_0)$ by the definition of $<$, and $A \in \mathbf{Desc}^{\mathcal{G}}(Z_0)$ by the definition of \mathcal{H} , and $V \in \mathbf{Desc}^{\mathcal{G}}(Z_0)$, since A is an ancestor of V .

Any fork in a path must either be an ancestor of the initial endpoint (in this case Z), or an ancestor of a collider in the path. Since $V \in \mathbf{Desc}^{\mathcal{G}}(Z_0)$ and V is a fork, not an endpoint, V cannot be an ancestor of the initial endpoint. So V must be an ancestor of a collider in the walk w . As w is active given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, the collider D must be in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$. We consider three sub-cases: 2.2.3.1: D is in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}] \setminus (\mathbf{X}' \cup \mathbf{C}')$, 2.2.3.2: D is in $\mathbf{X}'_{<C}$, 2.2.3.3: D is in $\mathbf{C}'_{<C}$, and will prove a contradiction in each case. *Sub-case 2.2.3.1: D is in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}] \setminus (\mathbf{X}' \cup \mathbf{C}')$.* Then all the parents of $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ must also be in $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$ by the definition of implied variables, and these parents would be non-colliders, which would make w blocked given $[(\mathbf{X}' \cup \mathbf{C}')_{<C}]$, giving a contradiction. *Sub-case 2.2.3.2: D is in $\mathbf{X}'_{<C}$.* Then at least one parent of D must be a non-collider in $\mathbf{C}'_{<C}$, which contradicts the statement that w contains no non-collider in $(\mathbf{X}' \cup \mathbf{C}')$. *Sub-case 2.2.3.3: D is in $\mathbf{C}'_{<C}$.* Then $D \in \mathbf{Desc}^{\mathcal{G}}(Z_0)$ (since $D \in \mathbf{Desc}^{\mathcal{G}}(V)$ and $V \in \mathbf{Desc}^{\mathcal{G}}(Z_0)$). It follows that $Z_0 < D$, but this contradicts the definition of C as the nearest variable along w to Z_0 that is in $\mathbf{C}'_{>Z_0}$.

From Lemma 30 the result follows. \square

References

- Scott Alexander. AI researchers on AI risk. *Slate Star Codex [blog]*, 2015. URL <http://slatestarcodex.com/2015/05/22/ai-researchers-on-ai-risk>.
- Benjamin Aminof, Marta Kwiatkowska, Bastien Maubert, Aniello Murano, and Sasha Rubin. Probabilistic strategy logic. *IJCAI*, 2019.
- Dario Amodei and Jack Clark. Faulty reward functions in the wild, 2016. URL <https://blog.openai.com/faulty-reward-functions>.
- Arbital. Problem of fully updated deference, 2017. URL https://arbital.com/p/updated_deference. Accessed: 2023-02-09.
- Stuart Armstrong. Utility indifference. Technical report, Technical Report 2010-1. Oxford: Future of Humanity Institute, 2010.
- Stuart Armstrong and Benjamin Levinstein. Low impact artificial intelligences. *arXiv preprint arXiv:1705.10720*, 2017.
- Stuart Armstrong and Xavier O’Rourke. ‘Indifference’ methods for managing agent rewards. *arXiv preprint arXiv:1712.06365*, 2017a.
- Stuart Armstrong and Xavier O’Rourke. Good and safe uses of ai oracles. *arXiv preprint arXiv:1711.05541*, 2017b.
- Stuart Armstrong, Jan Leike, Laurent Orseau, and Shane Legg. Pitfalls of learning a reward function online. *arXiv preprint arXiv:2004.13654*, 2020.
- Kenneth J Arrow. *Social choice and individual values*. Yale University Press, 1963.

- Hal Ashton. Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, pages 1–32, 2022.
- Hal Ashton and Matija Franklin. The problem of behaviour and preference manipulation in AI systems. *CEUR workshop proceedings*, 3087, 2022.
- Carolyn Ashurst, Ryan Carey, Silvia Chiappa, and Tom Everitt. Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. *IJCAI*, 2005.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022.
- Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 237–254. Association for Computing Machinery, 2022.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

- Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing AI risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, 2023.
- Sebastian Benthall and David Shekman. Designing fiduciary artificial intelligence. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–15, 2023.
- Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety—a review. *arXiv preprint arXiv:2404.14082*, 2024.
- Liad Blumrosen and Noam Nisan. Algorithmic game theory. *Introduction to Mechanism Design*, Cambridge University Press, New York, USA, 2007.
- Nick Bostrom. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and technology*, 9, 2002.
- Nick Bostrom. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3):308–314, 2003.
- Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.
- Nick Bostrom. Existential risk prevention as global priority. *Global Policy*, 4(1): 15–31, 2013.
- Nick Bostrom. *Superintelligence: Paths, dangers, strategies.*, 2014a. Oxford University Press, 2014.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

- Ryan Carey. In corrigibility in the CIRL framework. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 30–35, 2018.
- Ryan Carey and Tom Everitt. Human control: Definitions and algorithms. *Uncertainty in AI (UAI)*, 2023.
- Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.
- Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*, pages 2686–2708. PMLR, 2022.
- Center for AI Safety. Statement on AI risk, 2023. URL <https://www.safe.ai/statement-on-ai-risk>. Accessed: 2023-09-11.
- Krishnendu Chatterjee, Thomas A Henzinger, and Nir Piterman. Strategy logic. *Information and Computation*, 208(6):677–693, 2010.
- Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, 2019.
- Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.
- Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- Paul Christiano. Corrigibility, 2017. URL <https://ai-alignment.com/corrigibility-3039e668638>.

- Paul Christiano. What Failure Looks Like, 2019. URL <https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like>. Accessed 2023-04-28.
- Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, and Kate Smaje. The economic potential of generative AI. Technical report, McKinsey & Company, 2023.
- Iain M Cockburn, Rebecca Henderson, and Scott Stern. *The impact of artificial intelligence on innovation*, volume 24449. National Bureau of Economic Research Cambridge, MA, USA, 2018.
- Michael K. Cohen, Badri N. Vellambi, and Marcus Hutter. Asymptotically unambitious artificial general intelligence. In *AAAI Conference on Artificial Intelligence*, 2020.
- Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.
- Kate Crawford and Ryan Calo. There is a blind spot in AI research. *Nature*, 538 (7625):311–313, 2016.
- David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv preprint arXiv:2405.06624*, 2024.
- A Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 2002.
- Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

- Oren Etzioni. No, the experts don't think superintelligent AI is a threat to humanity. *Technology Review*, September, 2016.
- Tom Everitt, Ramana Kumar, Victoria Krakovna, and Shane Legg. Modeling AGI safety frameworks with causal influence diagrams. *arXiv preprint arXiv:1906.08663*, 2019a.
- Tom Everitt, Pedro A Ortega, Elizabeth Barnes, and Shane Legg. Understanding agent incentives using causal influence diagrams, part i: single action settings. *arXiv preprint arXiv:1902.09980*, 2019b.
- Tom Everitt, Ryan Carey, Eric Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI-21)*. Virtual. Forthcoming, 2021a.
- Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 2021b.
- Tom Everitt, Lewis Hammond, Francis R Ward, Ryan Carey, James Fox, Matt McDermott, Sebastian Benthall, and Jonathan Richens. Towards causal foundations of safe AGI. *Alignment Forum*, 2023. URL <https://www.alignmentforum.org/s/pcdHisDEGLbxrbSHD>.
- Enrico Fagioli and Marco Zaffalon. A note about redundancy in influence diagrams. *International Journal of Approximate Reasoning*, 1998.
- Sebastian Farquhar, Ryan Carey, and Tom Everitt. Path-specific objectives for safer agent incentives. *AAAI Conference on Artificial Intelligence*, 2022.
- Dana Fisman, Orna Kupferman, and Yoad Lustig. Rational synthesis. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 190–204. Springer, 2010.

- James Fox, Tom Everitt, Ryan Carey, Eric Langlois, Alessandro Abate, and Michael Wooldridge. Pycid: a python library for causal influence diagrams. In *Scientific Computing with Python Conference (SciPy)*, 2021.
- David Galles and Judea Pearl. Axioms of causal relevance. *Artif. Intell.*, 97(1-2): 9–43, 1997. doi: 10.1016/S0004-3702(97)00047-7. URL [https://doi.org/10.1016/S0004-3702\(97\)00047-7](https://doi.org/10.1016/S0004-3702(97)00047-7).
- Dan Geiger and Judea Pearl. On the Logic of Causal Models. *Machine Intelligence and Pattern Recognition*, 9:3–14, 1990.
- Tim Genewein, Tom McGrath, Grégoire Delétang, Vladimir Mikulik, Miljan Martic, Shane Legg, and Pedro A Ortega. Algorithms for causal reasoning in probability trees. *arXiv preprint arXiv:2010.12237*, 2020.
- Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*, 2024.
- David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa’s explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021.
- Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.
- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J Russell. The off-switch game. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 220–227, 2017.
- Joseph Halpern and Max Kleiman-Weiner. Towards formal definitions of blame-worthiness, intention, and moral responsibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

- Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, , Alessandro Abate, and Michael Wooldridge. Reasoning about causality in games. *AI Journal*, 2023.
- David Heckerman and Ross Shachter. A decision-based view of causality. In *Uncertainty Proceedings 1994*, pages 302–310. Elsevier, 1994.
- David Heckerman and Ross Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.
- Corinna Hertweck and Tim Rüz. Gradual (in) compatibility of fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- Koen Holtman. Corrigibility with utility preservation, 2020.
- Michael C Horowitz. When speed kills: Lethal autonomous weapon systems, deterrence and stability. In *Emerging Technologies and International Stability*, pages 144–168. Routledge, 2021.
- Ronald A Howard. *Decision analysis: Applied decision theory*. Stanford Research Institute, 1966a.
- Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966b.
- Ronald A Howard. From influence to relevance to knowledge. *Influence diagrams, belief nets and decision analysis*, pages 3–23, 1990.
- Ronald A Howard and James E Matheson. The principles and applications of decision analysis. *Strategic Decisions Group, Palo Alto, CA*, 1984.
- Ronald A Howard and James E Matheson. Influence diagrams. *Decision Analysis*, 2(3):127–143, 2005.
- Leonid Hurwicz. On informationally decentralized systems. *Decision and organization: A volume in Honor of J. Marschak*, 1972.

- Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, 2018.
- Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.
- Frederik Hytting Jørgensen, Sebastian Weichwald, and Jonas Peters. Unfair utilities and first steps towards improving them. *arXiv preprint arXiv:2306.00636*, 2023.
- Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *Artificial Intelligence*, page 103963, 2023.
- Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.
- Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*, 2023.
- Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B Tenenbaum. Inference of intention and permissibility in moral decision making. In *CogSci*, 2015.
- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
- Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.
- Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew

- Botvinick, and Christopher Summerfield. Human-centred mechanism design with democratic AI. *Nature Human Behaviour*, 6(10):1398–1407, 2022. doi: 10.1038/s41562-022-01383-x. URL <https://doi.org/10.1038/s41562-022-01383-x>.
- Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. *arXiv preprint arXiv:1806.01186*, 2018.
- Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19064–19074. Curran Associates, Inc., 2020a. URL <https://proceedings.neurips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf>.
- Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. *DeepMind Blog*, 3, 2020b.
- David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*, 2020.
- Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Advances in Neural Information Processing Systems (Neurips)*, 2017a.
- Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017b.
- Eric Langlois and 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. Forthcoming*, 2021.
- Steffen L Lauritzen and Dennis Nilsson. Representing and solving decision problems with limited information. *Management Science*, 47(9):1235–1251, 2001.

- Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in neural information processing systems*, 33, 2020.
- Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018.
- Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.
- Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 2020.
- Jarryd Martin, Tom Everitt, and Marcus Hutter. Death and suicide in universal artificial intelligence. In Bas Steunebrink, Pei Wang, and Ben Goertzel, editors, *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*, pages 23–32, Cham, 2016. Springer International Publishing. ISBN 978-3-319-41649-6.
- James E Matheson. Using influence diagrams to value information and control. *Influence diagrams, belief nets, and decision analysis*, pages 25–48, 1990.
- Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022.
- C Meek. Strong completeness and faithfulness in bayesian networks. unpublished, 1995.

- Allen C Miller III, Miley W Merkhofer, Ronald A Howard, James E Matheson, and Thomas R Rice. Development of automated aids for decision analysis. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1976.
- Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. Should robots be obedient? *IJCAI*, 2017.
- Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.
- Scott Mueller and Judea Pearl. Personalized decision making—a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.
- Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. PMLR, 2019.
- Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. In *Conference on Causal Learning and Reasoning*, pages 594–617. PMLR, 2022.
- Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.
- Thomas D Nielsen and Finn V Jensen. Welldefined decision scenarios. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 502–511. Morgan Kaufmann Publishers Inc., 1999.
- N Nisan, T Roughgarden, E Tardos, and VV Vazirani. Algorithmic game theory, cambridge univ, 2007.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

- Stephen M Omohundro. The basic AI drives. In *AGI*, 2008.
- Laurent Orseau. The multi-slot framework: A formal model for multiple, copiable AIs. In *Artificial General Intelligence*, volume 8598 LNAI, pages 97–108. Springer, 2014. ISBN 9783319092737.
- Laurent Orseau and Stuart Armstrong. Safely interruptible agents. *Conference on Uncertainty in Artificial Intelligence*, 2016.
- Ayodeji Oseni, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661*, 2021.
- Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pages 15–17, 1985.
- Judea Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.
- Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.
- Jonathan G Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. *arXiv preprint arXiv:2204.12993*, 2022.

- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- Stuart Russell. Human-compatible artificial intelligence. *Human-like machine intelligence*, pages 3–23, 2021.
- Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.
- Aaron L Sarvet and Mats J Stensrud. Perspectives on harm in personalized medicine. *arXiv preprint arXiv:2302.01371*, 2023.
- Aaron L Sarvet and Mats J Stensrud. Rejoinder to " perspectives onharm'in personalized medicine—an alternative perspective". *arXiv preprint arXiv:2403.14869*, 2024.
- Ross Shachter and David Heckerman. Pearl causality and the value of control. *Heuristics, Probability, and Causality: A Tribute to Judea Pearl*, pages 431–447, 2010.
- Ross D Shachter. Evaluating influence diagrams. *Operations research*, 34(6):871–882, 1986.
- Ross D Shachter. Bayes-Ball: The Rational Pastime (for Determining Irrelevance and Requisite Information in Belief Networks and Influence Diagrams). *Uncertainty in Artificial Intelligence (UAI)*, 1998.
- Ross D Shachter. Decisions and dependence in influence diagrams. In *Conference on Probabilistic Graphical Models*, pages 462–473. PMLR, 2016.
- Lee Sharkey, Clíodhna Ní Ghuidhir, Dan Braun, Jérémy Scheurer, Mikita Balesni, Lucius Bushnaq, Charlotte Stix, and Marius Hobbhahn. A causal framework for AI regulation and auditing. unpublished, 2024.

- Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.
- Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- Jessica Taylor. Two problems with causal-counterfactual utility indifference. *Alignment Forum*, 2016a.
- Jessica Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016b.
- Max Tegmark and Steve Omohundro. Provably safe systems: the only path to controllable AGI. *arXiv preprint arXiv:2309.01933*, 2023.
- Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.
- Jin Tian and Judea Pearl. Causal discovery from changes. *arXiv preprint arXiv:1301.2312*, 2013.
- Philip Trammell and Anton Korinek. Economic growth under transformative AI. Technical report, National Bureau of Economic Research, 2023.
- Alan M Turing. Can digital computers think? *Transcription from a BBC Radio Interview Cambridge*, 1951.
- Alexander M. Turner. Non-obstruction: A simple concept motivating corrigibility. *Alignment Forum*, 2020.

- Alexander M. Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020.
- Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. *NeurIPS*, 2021.
- Jonathan Uesato, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. Avoiding tampering incentives in deep rl via decoupled approval. *arXiv preprint arXiv:2011.08827*, 2020.
- Chris van Merwijk, Ryan Carey, and Tom Everitt. A complete criterion for value of information in soluble influence diagrams. *AAAI*, 2022.
- Thomas Verma and Judea Pearl. Causal Networks: Semantics and Expressiveness. In *Uncertainty in Artificial Intelligence (UAI)*, 1988.
- Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. Association for Computing Machinery (ACM), 2022.
- Francis Rhys Ward, Tom Everitt, Francesco Belardinelli, and Francesca Toni. Honesty is the best policy: Defining and mitigating AI deception. in submission, 2023. URL <https://causalincentives.com/pdfs/deception-ward-2023.pdf>.
- Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. The reasons that agents act: Intention and instrumental goals. *AAMAS*, 2024.
- Michael Webb. The impact of artificial intelligence on the labor market. *Available at SSRN 3482150*, 2019.
- Michael Wooldridge, Julian Gutierrez, Paul Harrenstein, Enrico Marchioni, Giuseppe Perelli, and Alexis Toumi. Rational verification: From model checking to

- equilibrium checking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- Sewall Wright. The theory of path coefficients a reply to niles’s criticism. *Genetics*, 8(3):239, 1923.
- Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in neural information processing systems*, 33:12263–12274, 2020.
- Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *International Joint Conference on Artificial Intelligence*, 2017.